Data Integration for the Analysis of Uncharacterized Proteins in *Mycobacterium tuberculosis*

Gaston KUZAMUNU MAZANDU

Thesis Presented for the Degree of
DOCTOR OF PHILOSOPHY
in Computational Biology
UNIVERSITY OF CAPE TOWN

Supervisor: Assoc. Prof. Nicola Mulder

August 2010

Declaration

I, the undersigned, hereby declare that the work contained in this thesis is my own original work and that I have not previously in its entirety or in part submitted it at any university for a degree.

Signature: ..... Date: .....

Abstract

Mycobacterium tuberculosis is a bacterial pathogen that causes tuberculosis, a leading cause of human death worldwide from infectious diseases, especially in Africa. Despite enormous advances achieved in recent years in controlling the disease, tuberculosis remains a public health challenge. The contribution of existing drugs is of immense value, but the deadly synergy of the disease with Human Immunodeficiency Virus (HIV) or Acquired Immunodeficiency Syndrome (AIDS) and the emergence of drug resistant strains are threatening to compromise gains in tuberculosis control. In fact, the development of active tuberculosis is the outcome of the delicate balance between bacterial virulence and host resistance, which constitute two distinct and independent components. Significant progress has been made in understanding the evolution of the bacterial pathogen and its interaction with the host. The end point of these efforts is the identification of virulence factors and drug targets within the bacterium in order to develop new drugs and vaccines for the eradication of the disease. However, a major limitation has been the high proportion of uncharacterized proteins encoded by the genome. In this thesis, we have integrated biological data from different sources for analyzing the unknown genes of *Mycobacterium tuberculosis* using the functional network produced.

Three main contributions to the field can be identified in this thesis. First we set up a data-driven scoring scheme for sequence and microarray data in order to fill gaps found in the existing and commonly used functional networks for this specific organism. Second we perform a functional analysis of the network produced through function prediction, where possible, of proteins labelled 'uncharacterized' in the organism using Gene Ontology (GO) terms. To this end, we set up a novel GO semantic similarity metric for comparing terms in the Gene Ontology structure for more efficient annotation prediction, thus contributing

to the curation of the proteome of this organism. Finally, we analyze the structure of the functional network to elucidate proteins essential to the functioning of the system, potentially contributing to the survival of the bacterium within the host and allowing the bacterial pathogen to prosper. These proteins are potential drug targets, and can be used to enhance the discovery process of new drugs in order to overcome this public health challenge. Drug targets have been traditionally identified through complete knowledge of individual proteins and their well characterized functions. Here, we integrate biological data from different sources into a single functional network to provide a systems view of the whole bacterial pathogen for the identification of new potential drug targets.

Acknowledgements

First I would like to express my deep recognition to my thesis supervisor Prof. Nicola J. Mulder who made this dissertation possible and introduced me to a wider academic Bioinformatics community. She has supported me through the difficult process of defining the area of research and later bringing those ideas to actual contributions. She has had an infinite patience to discuss with me on and on, even when my ideas were not clear sometimes. For her unlimited help and dedication through the whole process, for helping me beyond her duties at a human level, my sincere thanks.

I would like to thank my colleagues of the Computational Biology (CBIO) research group at the University of Cape Town for their encouragement and willingness to help, and for making my time enjoyable and stimulating during this process. A special thanks goes to Victoria Nembaware for reading the draft versions of each chapter and suggesting some improvement. I would like to extend my thanks to Bruno Le Floch for his time and availability to reread the whole document. In all the cases, the discussion about technical matters within and outside have contributed to the improvement of this thesis. Anyone who has passed by the research domain knows there is more to it than business.

May all those who have directly or indirectly contributed to the realization of this dissertation, find through this line the mark of my gratitude.

To my parents for their support through time and distance. I am proud of bringing them this achievement.

I am grateful to the trustee of the D. M. Mackay trust for the financial help, sympathy and compassion at the time when everything went dark with uncertain future. This has contributed to the completion of this thesis work. My special thanks goes also to Prince

Kaleme's family who has been supportive during this process.

But mostly I owe all this to my family, Malungidi Mbambi Marie-Paul my wife and our children. From the bottom of my heart, I would like to thank my wife Malungidi Mbambi Marie-Paul for her love and support during this seemingly never ending student life. She has tried her best to keep children on the 'good and straight line', playing mom and almost dad roles. Thanks to you, my champions Jemima Kuzamunu Mambote, Glodi Kuzamunu Mazandu, Keren Kuzamunu Kinzuemi, and Emmanuel Kuzamunu Malungidi for your patience and persistence at every step of the long way of this journey, and for making me one of the happiest fathers in the world. This increased my determination to keep fighting until I won a few. This dissertation is dedicated to you.

Any work dependent on open-source software owes debt to those who developed these tools. I thank everyone involved with free software, from the core developers to those who contributed to the documentation. Many thanks to the authors of the freely available libraries. This work was typeset with LaTeX on a computer running the GNU/Linux (Ubuntu) operating system where I frequently used gedit and gnuplot.

This work would not have been possible without financial support of National Research Foundation and National Bioinformatics Network (NRF-NBN) in South Africa through Computational Biology (CBIO) research group at the Department of Clinical Laboratory Sciences/Institute of Infectious Disease and Molecular Medicine (IIDMM), University of Cape Town.

List of Related Publications

[1] Gaston K. Mazandu, Kenneth Opap, and Nicola J. Mulder. *Contribution of Microarray Data to the Advancement of Knowledge on the Mycobacterium tuberculosis Interactome: use of the random Partial Least Squares approach.* Infection, Genetics and Evolution (IGE), Accepted September 2010 (`http://dx.doi.org/10.1016/j.meegid.2010.09.003`).

[2] Gaston K. Mazandu and Nicola J. Mulder. *Scoring Protein Relationships in Functional Interaction Networks Predicted from Sequence Data.* PLoS ONE, Accepted April 2011 (`http://dx.plos.org/10.1371/journal.pone.0018607`).

[3] Gaston K. Mazandu and Nicola J. Mulder. *GO-Universal Metric for Measuring Term Closeness in Gene Ontology (GO).* Under review 2011.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Throughout history, infectious diseases caused by microbial pathogens have been a scourge for mankind. Their devastating impacts on human morbidity and mortality remain of great concern, even today. With the advance of new high throughput sequencing technologies, there has been an increase in the number of worldwide microbial genome sequencing projects (`http://microbialgenome.org`, `http://www.ncbi.nlm.gov/genomes/MICROBES/Complete.html`, `http://www.sanger.ac.uk/Projects/Microbes` and `http://www.tigr.org/tdb/mdbcomplete.html`), which has yielded complete genome sequences of crucial microbial pathogens of humans, animals and plants. Analyses of these genome sequences have provided valuable insights into the dynamics driving pathogenic mechanisms and numerous virulence factors, and have shed light on the targeted organism's biology [2]. The characteristic features of pathogenic organisms include their ability to colonize a specific host organ or tissue, to adapt to their environment, and to evade the host immune response [3], thus leading to the development of disease, as a result of a delicate and dynamic balance between pathogen and host defence system.

The genome sequences allow the analysis of organisms, facilitating the identification of the genetic differences between several genomes from the same species, with the potential to generate information and knowledge about functional interactions between genes in an organism and genome relatedness [4]. Furthermore, the availability of these pathogenic microbial genomes can contribute to speeding up the process of drug target selection [5] by finding genes that are essential to microbial cell survival or growth and virulence. In fact, significant progress has been made in drug discovery and vaccine administration against

major infectious diseases [6]. However, these efforts are weakened by an increased incidence of widespread drug-resistant strains to the available and commonly used antibiotics and vaccines, a growing prevalence of infections, and the emergence of new pathogenic organisms, making infectious diseases the leading cause of human death worldwide. Tuberculosis (TB) constitutes the biggest component of these infectious diseases, which claimed 1.8 million victims in 2008 and there were estimates of 9.4 million new cases that year (3.6 million of whom are women), including 1.4 million cases among people living with Human Immunodeficiency Virus (HIV) or Acquired Immunodeficiency Syndrome (AIDS) according to the World Health Organization (WHO) [7, 8].

1.1 Pathogenesis and Virulence of MTB

TB is caused by an intracellular pathogen *Mycobacterium tuberculosis* (MTB), also known as tubercle or Koch's bacillus, whose genome sequence has been completely elucidated [1, 9, 10]. The progression of an MTB infection to disease is triggered by its virulence and pathogenesis, which mainly depend on the response of the host immune system and the size of the infecting dose of MTB [11]. Virulence is the mechanism by which microbes evade the host immune response [12] while pathogenesis is the ability of microbial pathogens to cause disease in their host. These processes are achieved through molecular interactions between specific microbial products and host cells [13]. This leads to modification of host cell functioning, thus allowing the parasite to invade host organs and tissues to ensure the parasite's survival, and resulting in disease in the host.

To control the growth of an MTB infection, the host engages a strong cell-mediated immune response, releasing interferons (IFNs) secreted by the CD4+ T lymphocytes to allow macrophages to overcome the parasite [14]. The activated macrophages secrete tumour necrosis factor alpha (TNF-α), which promotes and stabilizes granuloma formation and apoptosis (programmed cell death) [15]. Thus, the infection is cordoned off but the bacteria can survive extracellularly in a latent or dormant state [16], as granulomas are not sterilizing structures. These granulomas may heal or enlarge and shed bacteria into the blood stream and lymphatic system, or alternatively they may liquefy or form a cavity, resulting in uninhibited bacterial growth into airways, which facilitates aerosol transmission

of the bacterium [17].

Infection occurs by inhaling infected droplets released by an individual with untreated pulmonary TB via coughing, sneezing, talking or singing [18]. MTB is an airborne pathogen [19] in its spreading mechanism, and these infected droplets can remain in the air for a long period of time. They can be killed by ultraviolet light, including sunlight, but they can survive in a dark and unventilated environment for several hours. Thus, the chance of TB transmission depends on the number of MTB expelled into the air, the environment (light or humidity), the closeness of contact, the immune status of the host and the virulence of the bacteria.

On the MTB side, mechanisms are triggered to modulate the host response. These mechanisms include the down-regulation of antigen molecules [20], IFN-γ-mediated activation of macrophages [21], and host cell apoptosis [22]. In fact, the host immune response is initiated inside the phagosome and ends inside the phagolysosome [23]. To overcome the host immune response, MTB first thwarts the maturation of the phagolysosome by stopping the fusion of its phagosome with lysosomes [24]. A cell wall associated glycolipid, mannosylated lipoarabimannan (ManLAM) found in MTB attenuates expression of TNF-α and interferes with the recruitment of early endosome auto-antigen 1 (EEA1), a molecule involved in endosome-endosome fusion, leading to the failure of phagolysosome fusion [25]. It has also been shown that MTB uses multiple receptors including complement receptors 1 and 3 to break the border of host cells [26] and produces catalase (KatG) and super-oxide dismutase (SOD) enzymes to deteriorate reactive oxygen intermediates it is exposed to [27, 28]. The inactivation of the gene for catalase peroxidase, KatG, may lead to the attenuation of growth of the bacillus [2]. Thus, understanding the molecular basis of mycobacterial virulence depends on the identification of genes and gene products that contribute to the pathogenesis, and the challenge is to optimally use available biological data for accelerating the identification of new virulence factors.

In TB, the bacteria mainly affect the lungs, where they are initially ingested by pulmonary macrophages and undergo intracellular multiplication, and where infection starts. They can switch from an active to dormant state, called a latent state, in which case the infected person does not suffer from TB. In the 1890s, Robert Koch demonstrated that the tubercle bacillus does not produce toxins [2], meaning that the pathogenicity associated with the

disease mainly results from excessive cell-mediated immune and inflammatory responses mounted by the host in response to bacterial antigens. Therefore, disease occurs when the host immune system is compromised or becomes weak, allowing the bacteria to successfully overcome its immune defence, multiply and become numerous enough to cause damage to the lungs. The infection spreads to other parts of the body or other organs [18], such as the brain, spine, or kidneys via blood circulation. There are two main forms of TB [29, 30], namely extra-pulmonary tuberculosis (EPTB), which refers to disease outside the lungs, and the one attacking the lungs, referred to as pulmonary TB (PTB). The contagious factors depend on the form of tuberculosis, but in general, only the pulmonary form of infections is contagious.

1.2 Tuberculosis Epidemiology, Treatment and AIDS

In 1882, a German bacteriologist Robert Koch established the aetiology of TB, and stated [31]: "If the importance of a disease for mankind is measured by the number of fatalities it causes, then tuberculosis must be considered much more important than other most feared infectious diseases, plague, cholera and the like. If one only considers the productive middle-age groups, tuberculosis carries away one-third, and often more". This statement still holds word for word after more than one and a quarter centuries, with no hope so far of changing. Indeed, according to the World Health Organization (WHO) [7, 8], tuberculosis remains a potent infectious killer, as more than two billion people, corresponding to approximately one-third of the world's population, are infected with MTB, of which, one in ten develops active tuberculosis, affecting mostly young adults in their most productive years. The objective is to reduce the incidence of TB by half by 2015, thus hard work is required if we are to achieve the Millennium development Goal of halting and beginning to reverse the spread of TB as one of the world's major diseases [32].

The first and widely used BCG (Bacille Calmette-Guérin) vaccine against TB was discovered in 1906 by the two French biologists, Albert Calmette and Camille Guérin. It was used for the first time in 1921. In addition, several compounds were introduced between 1943 and 1963 as anti-TB agents [33]. These include streptomycin (1943), p-aminosalicylic acid (1949), isoniazid (1952), pyrazinamide (1954), cycloserine (1955), ethambutol (1962),

and rifampin (1963). This rapid succession of drugs with anti-TB activity was aimed at accelerating the exit of TB as a public health challenge. Today, over 20 drugs against TB are available [34] but TB is a seemingly unstoppable disease and a leading cause of human death from infectious diseases. This clearly demonstrates the inefficiency of global TB eradication programmes and the ineffectiveness of tuberculosis control measures and treatment protocols implemented so far. The widespread emergence of drug resistant strains constitutes a major impediment to these global TB eradication programmes, requiring the effectiveness of coordinated strategies towards new anti-TB compounds with novel mechanisms of action.

One of the factors determining the virulence of MTB is its ability to switch from a replicative (growth) to a non-replicative (dormancy) state [35], in response to the antimycobacterial host defence mechanisms. This enables it to persist, infect, grow and survive in human macrophages, thus making this slow growing aerobic bacterium, one of the most successful human bacterial pathogens. Therefore, the properties of anti-TB agents must include antibacterial activity, capacity to inhibit the development of resistance, and capacity to kill these intracellular, persisting organisms. For this purpose, the initial combination of isoniazid, rifampin, pyrazinamide, and ethambutal are used as front-line drugs [36]. Other anti-TB agents, namely aminoglycosides/cyclic peptide such as capreomycin, viomycin, kanamycin and amikacin, and quinolones such as moxifloxacin, levofloxin, ofloxacin and ciprofloxacin are effective but usually used in drug resistance situations [37, 38].

Multidrug-resistant TB (MDR-TB) is a form of TB that is resistant to at least two of the most commonly used drugs in the current four-drug or first-line therapies (isoniazed and rifampin). Extensively drug-resistant TB (XDR-TB) occurs when there is resistance to any fluoroquinolone, and at least one of the three injectable second-line drugs (capreomycin, kanamycin, and amikacin) on top of MDR-TB [39]. MDR-TB treatment is difficult and expensive [8], as much as 1400 times the cost of that of the regular treatment, and often unaffordable by persons that are infected. The global HIV epidemic has led to an explosive increase in TB incidence and contributes to increases in MDR-TB prevalence [40]. There were estimates of 500 000 new MDR-TB cases, 9.27 million incidences and 13.7 million prevalent cases of TB in 2007 [7, 8]. Specifically, Sub-Saharan Africa has all of the conditions of a 'perfect storm' for HIV infection and MDR-TB [41].

A weakening in the immunological defence system of the host, resulting in dissemination of the bacterial pathogen underlies the development of TB. Thus, much ground in the fight against TB has been lost, especially in the face of the spreading of HIV/AIDS, counted among leading infections compromising the host immune system. In Sub-Saharan Africa, for example, approximately 75% of people with active TB are infected with HIV [8], in which case, TB often constitutes a death sentence [42]. TB is a leading killer of people with HIV and people living with HIV and infected with TB bacilli are 20 to 40 times more likely to become sick with active TB than people not infected with HIV living in the same country [43]. In 2008, 1.8 million people died from TB, including 500 000 people with HIV corresponding to 4500 deaths a day [8]. At least one-third of the 33.2 million people living with HIV worldwide are infected with TB and one in four people with HIV die due to TB [43].

1.3 Overview of the MTB Genome

The biggest step towards understanding MTB virulence and its specific abilities for invasion and division inside host macrophages and defeating the antibacterial mechanisms of these cells, was realized with the complete elucidation and publication of the first MTB genome sequence of H37Rv, the laboratory strain, in 1999 [9]. This has facilitated the identification of all MTB proteins and, using computational methods through comparative and functional genomics, the availability of these entire genomes has facilitated the identification of genes common to all bacteria or specific to MTB, and the prediction of missing functions for proteins, respectively. Even though the large number of uncharacterized genes limits these genomic studies, such data has provided a basis for selecting potential drug targets from the complete list of proteins. The genes in the MTB genome, but missing from closely related genomes, are likely to be crucial to its pathogenicity and constitute promising candidates for drug targets [5].

The global analysis of bacterial proteomes [44, 45] has yielded insights into functional features of several uncharacterized proteins. This has led to the elucidation of biological properties unique to MTB, such as its virulence, its slow growth and persistence, and the complexity of its cell wall, thus answering certain pressing and interesting questions

about the pathogenicity of the tubercle bacillus. The H37Rv or laboratory strain genome sequencing project was accomplished by the Sanger Centre (`http://www.sanger.ac.uk`), in collaboration with the Pasteur Institute in Paris. One isolated clinical strain of MTB, CDC1551, which is seen as highly transmissible and virulent for humans, was sequenced at the institute of Genomic Research (TIGR) (`http://cmr.jcvi.org/tigr-scripts/CMR/GenomePage.cgi?org=gmt`) in 2002 [10]. This offered the opportunity to compare the two genomes and several characteristics of these genomes that were previously unknown have been revealed [4]. In fact, such comparative analyses have provided several effective approaches for enhancing our understanding of the molecular basis of tubercle bacillus virulence, evolution and pathogenesis, and have made possible the analysis of intra-species variability of this slow-growing aerobic pathogenic bacteria.

Over the years, more genome information has become available and these data are stored in databases that are freely accessible via web interfaces (`http://genolist.pasteur.fr/Tuberculist`, `http://cmr.jcvi.org/tigr-scripts/CMR/shared/Genomes.cgi`, etc.). This has resulted in improved precision of functional assignments for genes and their products, and information has become more accurate. Certain gaps in the understanding of MTB biology have been filled, and the initial annotation of the MTB laboratory strain has become outdated. Four sequencing errors have been detected, changing its size from $4,411,529$ to $4,411,532$ bp [46]. We describe the MTB genome features in the following section, and survey strain comparisons in sections 1.3.2. The three strains considered are the clinical Oshkosh, or CDC1551 strain, ATCC25618 or H37Rv and ATCC25177 or H37Ra. The term 'strain' refers to a collection of organisms closely related genetically and used in the laboratory or with a documented pathology history.

1.3.1 Biological Organization of the Genome

The elucidation and analysis of the complete nucleotide sequences have revealed several features of the genomes of the MTB species presented in table 1.1 extracted from `http://cmr.jcvi.org/tigr-scripts/CMR/shared/Genomes.cgi`. These species of MTB have genomes with 65.6% G+C content, approximately 4×10^6 base pairs, and 4×10^3 open reading frames (ORFs). The annotation of these 4000 genes has been classified by Tuberculist into 11 functional categories as described in table 1.2 (Data provided for MTB

Table 1.1: *Features of the MTB genome.*

Strains	Features					
	Genome size (bp)	Protein Coding genes	GC content(%)	CDS coverage	rRNA	tRNA
H37Ra	4,419,977	4034	65.61%	90.84%	2	45
H37Rv	4,411,532	3948	65.61%	90.93%	5	45
CDC1551	4,403,837	4246	65.60%	92.42%	3	45

Table 1.2: *Distribution of MTB proteins per functional class.*

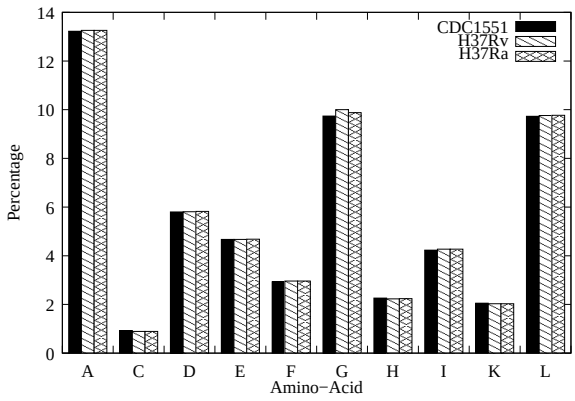
	Functional Class	Number of Proteins
1	Virulence, detoxification and adaptation	209
2	Lipid Metabolism	236
3	Information Pathways	232
4	Cell-wall and Cell Process	749
5	Stable RNAs	-
6	Insertion Sequences and Phages	109
7	PE/PPE/PGRS Proteins	167
8	Intermediary Metabolism and Respiration	895
9	Protein of Unknown Function	17
10	Regulatory Proteins	193
11	Conserved Hypothetical Proteins	1141

strain H37Rv from `http://genolist.pasteur.fr/Tuberculist`).

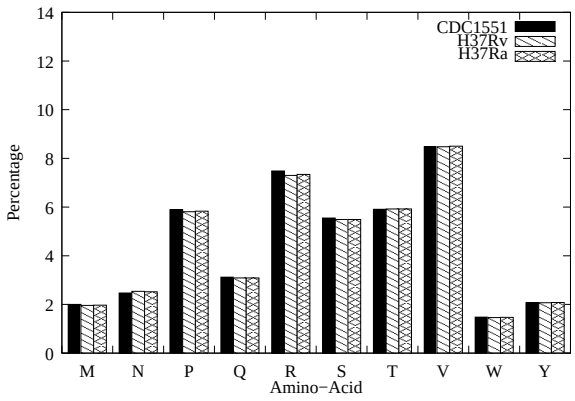
Predicted biological roles can be assigned to 66% of the ORFs, while 17 proteins have high sequence similarity with hypothetical proteins from other species and about 33% are regarded as novel proteins due to no match to any known ORFs in the current databases. The frequencies of amino acids in the whole MTB proteome are shown in figure 1.1.

Lipid Metabolism Genes

The cell envelope of MTB has been shown to contain a remarkable quantity of lipids [9], its genome sequence has revealed several genes dedicated to their production. Indeed, as shown in table 1.2, approximately 10% of proteins with known functional categories in the MTB genome are devoted to this activity, with genes involved in both biosynthetic and lipolytic pathways [46]. Several genes code for enzymes that catalyze reactions for



(a) Percentage of amino acids from Alanine (A) to Leucine (L)



(b) Percentage of amino acids from Methionine (M) to Tyrosine (Y)

Figure 1.1: *Distribution of amino acids in the MTB proteome.*

alternative sources of carbohydrates and are involved in alternative lipid degradation from the host vacuolar or cellular membranes, and hence contribute to energy metabolism [2]. It is essential for the survival of MTB in its host to have enzymes involved in lipid biosynthesis, as its envelope, which serves as an interface with the host, needs lipids, glycolipids, lipoglycans, and polyketides [47]. For example, a fused gene (pk1/15) encoding a phenolic glycolipid is involved in a decrease in pro-inflammatory cytokine production from host immune cells [48] and may contribute to MTB virulence [49].

The lipid components of the bacterial cell wall, such as lipoarabinomannan and sulfolipids, have been shown to modulate the immune response and protect the organism from the host defence system. Furthermore, MTB possesses a large number of transmembrane proteins named MmpL located next to genes involved in lipid metabolism, and observed to act as proton motive force-dependent efflux systems and confer high resistance to fluoroquinolones and other antibacterial agents in the gram negative bacteria *Pseudomonas aeruginosa* and *Escherichia coli* [2]. It has been suggested that these proteins are involved in the export of lipids or glycolipids and may act in drug efflux and hence contribute to the extensive natural resistance of the tubercle bacillus. It is not excluded that many other lipid metabolism proteins have features related to virulence, for example, the bacillus contains many genes involved in polyketide synthesis that may play a role in virulence, either as toxins or as immune modulators functioning like the macrolactone immunosuppressor rapamycin. Thus, an understanding of the composition and metabolism of these lipids may advance

our knowledge on the pathogenesis of MTB.

PE and PPE Protein Family

The PE and PPE protein families constitute one of the most important discoveries gained from MTB genome sequencing projects. These two novel large and unrelated families of acidic, glycine-rich proteins contain about 100 and 67 members [9], possessing a highly conserved N-terminal domain of about 110 and 180 amino acid residues, respectively. These sequences have characteristic motifs Pro-Glu at positions $8-9$ and Pro-Pro-Glu at $8-10$, where Pro and Glu stand for Proline (P) and Glutamic (E) amino acids, respectively, and followed by C-terminal segments that vary in sequence and length [2]. Each of these families falls into several subfamilies based on their C-terminal domains, the PE family is divided into three subfamilies, the most important is Polymorphic GC-Rich Sequences (PGRS), encoding the motif AsnGlyGlyAlaGlyGlyAla. On the other hand, PPE is composed of at least three classes of which the Major Polymorphic Tandem Repeats (MPTR) class constitutes the biggest component, encoding Asn-X-Gly-X-Gly-Asn-X-Gly [50].

The PE and PPE proteins are generally uncharacterized and their subcellular location is unknown except for Rv3097, which has been demonstrated to function as a lipase [9]. They form a source of antigenic variation among different strains of MTB [51] and might interfere with immune responses by inhibiting antigen processing [52]. Some predictions of these proteins indicate that they are expressed based on the changing micro-environments encountered by the pathogen and play an important role in survival and multiplication of MTB in their chosen environment, and even in mediating mycobacterium-host cell interactions [53, 54, 55]. This reveals that these proteins may provide to the pathogen the ability to switch from one metabolic path to another [56, 57] including aerobic and anaerobic respiration, thus allowing the organism to survive within its host in different environments ranging from high oxygen potential in the lungs to micro-aerobic/anaerobic conditions within the tuberculous granuloma. Furthermore, these protein families have been hypothesized to undergo recombination events, which increases their antigenic variability and may have profound implications in pathogenicity and/or host adaptation [58]. The fact that PE-PGRS proteins have been found to be exclusive to MTB complex organisms suggests that they are responsible for survival of the organism in host macrophages [59]. In addition, variations in the PE/PPE gene families might constitute a key to predominant differences

in pathogenesis between mycobacterial species [4]. Note, the term 'MTB complex' refers to a group of *Mycobacterium* species that cause TB in humans and other mammals [60].

ESAT-6 Protein Family

The 6-kDa Early Secretory Antigenic Target (ESAT-6) family, certain members of which are counted among eight ORFs in the Region of Difference 1 (RD1), has been shown to act as potent stimulators of the immune system. These members are assigned a cytolytic role [61], which enables virulent mycobacteria to spread, thus causing greater tissue damage [62]. The elucidation and annotation of the MTB genome revealed that 23 genes belong to ESAT-6 family and these genes are distributed in 11 different regions [9]. This family has been considered to be virulence factors as they seem to be important targets for T-cell responses [63].

Indeed, it has been shown that the removal of RD1 encoding ESAT-6 in H37Rv attenuated its virulence to that of the vaccine strain *Mycobacterium bovis* BCG [64]. According to Koch's postulate, this classifies the ESAT-6 protein family as virulence factors. This postulate stipulates that a gene is considered to be a virulence factor if the phenotype is found in virulent strains and its knock-out leads to attenuation, and that complementation of the mutant with the wide-type restores virulence [2]. Note that several other MTB antigens have also been identified, including Ag85A, the phosphate-binding protein PstS, the 36-kDA protein, and the heat-shock proteins hsp 60 and hsp 70 [65].

1.3.2 Strain Variation and Comparative Genomics

The well studied MTB strains, namely the laboratory strains H37Rv (virulent) and H37Ra (attenuated) both share a common ancestor, strain H37, derived from the same parental strain as the clinical strain CDC1551 [66], as described in figure 1.2.

Comparative genomics approaches have provided several insights into MTB strain evolution and pathogenesis. These include the discovery of 16 genomic regions, known as regions of difference (RD) $1-16$, deleted or missing in certain strains of *Mycobacterium bovis* BCG. Among these, 11 RDs, namely RD1, RD4-RD7, RD9-RD13 and RD15, contain 89 proteins of MTB strain H37Rv that are not found in any *Mycobacterium Bovis* BCG substrains

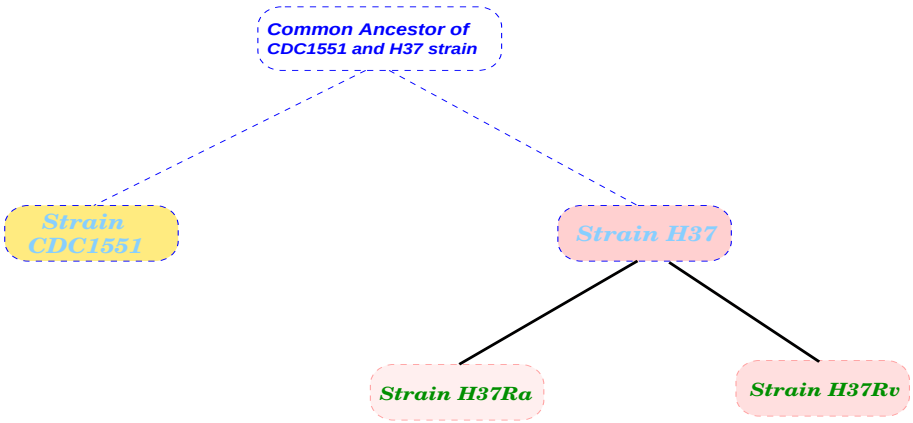


Figure 1.2: *Evolution of three MTB strains, H37Ra, H37Rv and CDC1551, adapted from [1].*

currently used as vaccines against TB worldwide [67]. In addition, six genomic deletion regions (RvD1-RvD6) [68] are missing in the MTB strain H37Rv relative to *Mycobacterium bovis*, but only 4 are absent in H37Ra, except for RvD2 and RvD6 [1]. In fact, before the elucidation of the genome of the MTB attenuated strain H37Ra, the molecular basis of attenuation of virulence in this strain was unknown, and the presence of these regions could not be used to explain this attenuation for they were present in clinical isolate (CDC1551).

Finally, about 4% of proteins with known functional categories in the MTB genome are comprised of mobile genetic elements (insertion sequences and prophages) [46] involved in genome plasticity of mycobacteria [69]. It has been shown that IS6110-related insertions or deletions constitute a key player in mediating genomic rearrangements and deletions in mycobacteria [1]. Four copies of IS6110 are found in CDC1551 compared to 16 copies in H37Rv and 17 in H37Ra [1, 10]. As these genetic variations among MTB strains might be a sign of selective pressure, they may play an important role in bacterial pathogenesis and immunity [10]. Some of the genes in the PE/PPE/PE-PGRS family in the attenuated strain are altered [1] under the influence of host immune selective pressure, likely allowing the bacteria to adapt to its environment during infection or transmission [70]. It has been suggested that these variations together with the lack of the expression of genes in the RD1 region, including ESAT-6 and 10-kDa culture filtrate protein (CFP-10) [71], may be implicated in the virulence attenuation of H37Ra [1].

In summary, this section surveys the available genomic and functional information on the

MTB genome. This enables the identification of genes and proteins which define the intrinsic features of MTB, thus providing a general view about MTB virulence, pathogenicity and interaction with the host immune system. Exploiting information from comparative genomics, several proteins have been suspected to play important roles in adhesion of the microbial pathogen in the host system and also in its intracellular lifestyle. These include protein members of the PE/PPE, lipid metabolism and virulence functional categories, and with them, many proteins of the unknown functional category. These data, not only provide a general view of the MTB genome and enhance our understanding of this organism, but also show that the use of available data and computational methods may help us better understand the mechanisms of virulence of MTB and features that enable this organism to adapt to or evade the host immune response.

1.4 Thesis Rationale

Despite the wide variety of anti-tuberculosis drugs, tuberculosis remains a leading cause of human death from infectious diseases [7, 34]. The most commonly used vaccine worldwide, BCG [67], offers some protection against TB infection in children but is considered to have a limited protective value in adults [72]. Even for infants, it prevents the spread of MTB within the body, but it does not prevent initial infection [73, 74]. With the emergence of drug resistant strains of MTB, the effectiveness of current anti-tuberculosis compounds is questionable, leading to an intensive and continuous search for new drugs with novel therapeutic agents. One of the biggest barriers to this process is the high proportion of proteins of unknown function in the MTB genome. Since the elucidation of the whole genome sequence and with the help of new high throughput biology technologies, significant successes have been recorded from comparative and functional genomics in gaining a better understanding of its evolution and interaction with its host. However, existing information on the microbial pathogen is still limited and incomplete. These include:

1. A significant number of hypothetical proteins in microbial genomes are labelled 'unknown' and currently uncharacterized in the protein databases.

2. Incomplete information on the functional interactions between proteins due to the nature and limitations of the methods used to derive them.

1.4.1 MTB Protein Annotation

The annotations of proteins in the MTB proteome were generally inferred from sequence homology searches, where functions of proteins of known function are assigned to a query protein using sequence similarity search tools, such as the Basic Local Alignment Search Tool (BLAST) [75]. This offers an easy and effective scheme of suggesting possible functions for proteins under consideration, but its applicability is limited. For instance, no known sequence may be similar to the query protein in current database. Moreover, duplication of genes after a speciation event can lead to inaccurate inference of function between homologous sequences where a duplicate of the original gene changes its function to an unrelated function in response to selective pressure. These annotations are not always manually verified. Therefore, though straightforward, this approach is limited and has thus left almost half of the proteins in the MTB proteome uncharacterized.

From the point of view of drug design, TB is currently treated with a decades-old drug regimen, lasting at least six months [29, 76] with no guarantee of the complete sterilization of the infection. This is, in part, due to the fact that the discovery of antibiotic drugs to treat the disease was based on the same therapeutic targets, and this strategy has failed to deliver a sufficient molecular diversity of drugs in order to overcome this public health challenge. The long and complex treatment of the disease is also leading to the development of resistance, and the treatment of drug resistant TB is even longer and more expensive than the regular treatment. The treatment of drug resistant strains may take between 18 and 24 months [73, 77], is often toxic and leads to negative side-effects, making patient compliance difficult. This constitutes one of the biggest limitations to any attempts in using derivatives of the existing anti-tuberculosis drugs for the development of new ones. Though attractive and economical, since the identification of novel drug targets for diseases and development of new drugs are always expensive and time-consuming, this strategy is limited for this specific organism.

In addition to hampering the search for new drug targets, progress towards the advancement of research on MTB and enhancement of our understanding of its virulence and pathogenicity is weakened by the high proportion of proteins of unknown function in its genome. Therefore, there is an urgent need for predicted functional annotations for this large number of uncharacterized proteins. This has the downstream potential to enable

researchers to apply these data to the search for new drug targets and thus for the development of novel and effective drugs with new biological mechanisms of action against drug susceptible and drug-resistant strains, reliably administered with a shorter regimen.

1.4.2 Existing MTB Functional Networks

Most processes in a living cell are accomplished through protein-protein interaction networks, therefore these play a central role in most activities involving the structure and function of the cell. These include signal transduction, protein folding, cell cycle control, DNA replication and transport [78]. The most commonly used integrated functional networks for many other organisms [34, 79, 80], are obtained from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database [81, 82]. In this functional network, edges are protein functional interaction links. These links are based on different evidence types, including direct or physical protein-protein interactions and many other types of possible functional connections. The STRING scoring system for protein or gene interactions is benchmarked by the Kyoto Encyclopedia for Genes and Genomes (KEGG) database [83] in which only 1028 out of more than 4×10^3 encoded proteins in the MTB proteome have a known pathway, representing about 25%. This constitutes the biggest limitation for scoring newly discovered interactions between genes and/or proteins, specifically for MTB which is not a model organism. In addition, the experimental data in the STRING database for this particular organism is very limited.

As an illustration, when dealing with microarray data, the STRING database retrieves its co-expression interactions from ArrayProspector (`www.bork.embl.de/ArrayProspector`) [84]. However, a large amount of microarray data for MTB are being generated and are publicly available in other resources, and these may increase the accuracy and precision of STRING data. In the case of homology data, the STRING scoring system relies on the $E - value$ obtained from sequence similarity searches. However, there are also protein signature databases such as InterPro [85], which is an integrated database for protein families and domains (`http://www.ebi.ac.uk/interpro`) [86], and can be used to increase the reliability and coverage of these homology data. Therefore, there is a need for an effective scoring system that does not rely on KEGG in order to fill gaps found in homology and microarray data in STRING for this specific organism in order to produce an MTB functional network

which constitutes the dynamic closure of the currently used MTB functional networks. This means that this MTB network can easily be reconstructed the instant the biological data for the organism are updated. We need to take advantage of different available data for MTB and integrate these, combining updated information from multiple interaction data sources into one unified network, with higher confidence and increased coverage.

This dynamic closure MTB functional network will enable researchers to better understand the biological processes that proteins are involved in. The interplay between proteins as well as each protein's role or function can be studied through the stability of the system under unchanging environmental conditions and the robustness of the system under changing environmental conditions or stressful perturbations.

1.5 Project Outline

The quantity of biological data is increasing exponentially resulting from worldwide DNA sequencing efforts and high-throughput biology technologies. We assume that these data can be categorized into two groups, namely primary data, such as genomic sequences, and secondary, or functional data. Building upon this assumption, this thesis revisits the problem of existing functional protein-protein networks to propose:

- New scoring systems for functional relationships extracted from sequence and microarray data.
- Data integration strategies which can be deployed for extensive integration of biological data into a single network with a specific focus here on the MTB genome.

The resulting network provides a mechanism for protein function prediction and gaining a better understanding of the biological organization of this organism. The main objectives of the project are thus two-fold:

1. Construction of the MTB functional network and use for protein function prediction. We have carried out large-scale integration of STRING data together with interactions derived from homology and microarray data produced using a dynamic

data-driven scoring system, yielding a dynamic closure of the existing protein-protein functional network, and enabling its update at any instant data are updated. This network is used to predict, where possible, functions of uncharacterized proteins on the basis of the Gene Ontology (GO) [87] functional annotations of their neighbours using the state-of-art similarity metric derived from a GO-universal metric. The existing guilt-by-association approaches require the exact match of terms used to characterize proteins when predicting functions of unknown proteins. This is not convenient when dealing with terms in the GO Directed Acyclic Graph (DAG), in which terms can be similar at a certain level without being identical. Therefore, we developed the GO-universal metric to overcome this limitation. We then use GO terms to predict, where possible, the functions of proteins labelled 'uncharacterized' by observing the pattern of their neighbours in the MTB functional network. In order to achieve better trade-off between improvement of quality, genomic coverage and scalability, the "Guilt-by-Association" approach has been extended to annotate these proteins by observing the key principles driving the functions imposed on a protein by its neighbours, referred to as 'traces' of underlying biological mechanisms.

2. Structure analysis of the network. The protein-protein functional network approach is being used more and more in the post-genomic era for a better understanding of cell functioning and organism development [34, 79]. We have conducted extensive computational analyses to dissect the MTB functional network produced in order to reveal the biological organization of the organism and to infer the role of the organism's proteins on the basis of the network's topological properties. This improves our understanding of protein functions and interactions and facilitates the identification of potentially important proteins that may be suitable drug targets, thus providing the opportunity to expand the range of potential drug targets and to move towards optimal target-based strategies.

These objectives are depicted in figure 1.3 which describes the approach used in the project from data integration to the main analyses conducted. In addition, data produced have been stored in a MySQL database and made accessible via a web interface at `http://lab12.cbio.uct.ac.za/tbannotations_v2/`.

The rest of this thesis is organized as follows: Chapter 2 addresses issues related to the current scoring approaches used for homology and microarray data, and introduces novel scoring schemes; an information-theoretic approach for homology data producing a con-

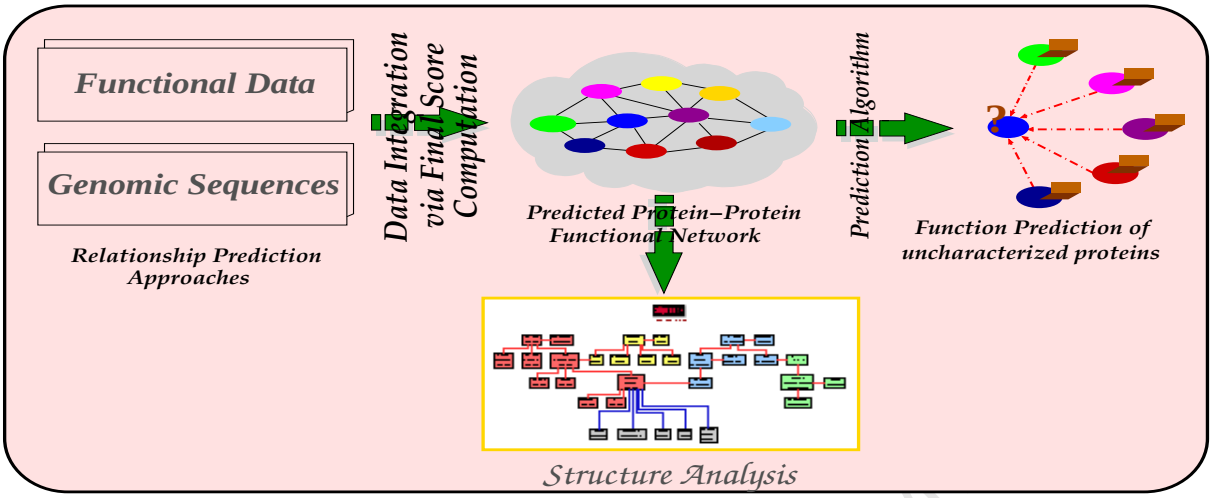


Figure 1.3: *Schematic representation of MTB genome analysis.*

fident homology network, and a partial least squares approach for microarray data, producing an MTB co-expression network with the exact complexity of the model, i.e., with the model that avoids risks of "over-fitting" or "under-fitting" data under consideration by controlling model parameters. In chapter 3, we infer the MTB functional network by combining heterogeneous data for protein-protein functional interactions and defining a unified weighting scheme in order to normalize or standardize these different schemes. In chapter 4, we use the network produced to perform functional analysis of the MTB proteome and predict, where possible, functions of uncharacterized proteins. The work presented in chapter 5 consists of using the functional network induced for identification and further analysis of potentially important genes on the basis of the network structure. We conclude the thesis in chapter 6, discussing future research directions.

Chapter 2

Integrative Scoring System for Sequence and Microarray Data

This project aims to build a functional association network for proteins in MTB to gain a better understanding of its biology. As mentioned above, existing resources like STRING have limitations in some of their data or methods, and here we address these for sequence homology and microarray information by incorporating additional data and developing new scoring methods for converting the data into pair-wise protein functional associations.

The abundance of diverse biological data from various sources constitutes a rich resource of genome knowledge from which functional association datasets can be derived. For example, from genome or protein sequences we can perform sequence similarity searches using the Basic Local Alignment Search Tool (BLAST) [75] in order to infer homology-based functional associations. Using data from protein domain and family databases, the inference of functional associations can be carried out based on the fact that two proteins sharing common domains or belonging to the same family are more likely to be functionally linked. These functional associations tend to be in Boolean or binary format, *i.e.*, either two genes or proteins are functionally linked, in which case the score is 1, or they are not and the score is 0. Such a scoring scheme for measuring functional links from different data types is not consistent or effective, since it does not take into account certain parameters, such as uncertainty of data, or noise inherent in the experiments used to derive these data.

Considering that we are dealing with inaccurate data obtained from different experi-

ments [88, 89], the uncertainty of data and noise inherent in each experiment must be efficiently managed by systematically weighing or scoring functional associations [81] derived from them. This is referred to as a reliability or confidence score of functional associations for the computational approach used for the prediction. These scoring schemes must be data source and technology dependent. This means that a given scoring scheme should normally vary according to the data sources and be designed on the basis of knowledge about the technology used. Furthermore, the effectiveness of a scoring scheme for functional associations is critical for the quality of post-analyses of the network, including functional and structural analysis, and will influence the predictions performed on the basis of these networks. Therefore, failure in setting up an appropriate scheme may negatively impact on the prediction analyses performed on the basis of these networks. This chapter revisits the existing scoring schemes of functional associations for data from protein domain and family databases, sequence similarity and microarray experiments to propose new efficient and effective schemes.

2.1 Scoring Inferences from Sequence Data

Scoring functional relationships from sequence data, which include protein family and shared domain, as well as sequence similarity data, has been widely addressed by the Bioinformatics community. However, the approaches used so far in the literature are limited to finding the similarity scores between proteins. In the case of protein family and domain data, this similarity score represents the number of common signatures shared by proteins (a protein signature is a model describing protein domains, families or sites). Two examples of such a scheme are given below.

Scheme 1: Scoring Function of Pfam Domain Sharing [89].
The scoring function $\mathcal{S}_{\text{pfam}}$ of Pfam domain sharing is simply the number of common domains shared, and given by

$$\mathcal{S}_{\text{pfam}}\left(\mathcal{P}_i, \mathcal{P}_j\right) = \left|\mathcal{D}_{\mathcal{P}_i} \cap \mathcal{D}_{\mathcal{P}_j}\right| \tag{2.1}$$

where $\mathcal{D}_{\mathcal{P}_k}$ is the set of Pfam domains found in protein $\mathcal{P}_k$.

Scheme 2: Scoring Function based on Protein Signature Profiling [90].
The similarity score between a pair of proteins $(\mathcal{P}_\imath, \mathcal{P}_\jmath)$ is computed using a binary similarity function between a pair of their signature profiles and given by

$$\mu\left(\mathcal{P}_\imath, \mathcal{P}_\jmath\right) = \frac{\sum_{\ell=1}^{n} \left(\mathbf{P}_\imath \wedge \mathbf{P}_\jmath\right)_\ell}{\sum_{\ell=1}^{n} \left(\mathbf{P}_\imath \vee \mathbf{P}_\jmath\right)_\ell} \tag{2.2}$$

where n is the number of signatures contained in proteins of a genome of interest and $\mathbf{P}_\ell = [S_{\ell 1}, S_{\ell 2}, \ldots, S_{\ell n}]$ the signature profile of protein $\mathcal{P}_\ell$, with $S_{\ell k} = 1$ if the signature $\mathcal{S}_k$ exists in protein $\mathcal{P}_\ell$ and $S_{\ell k} = 0$ otherwise.

For sequence similarity data, the scoring schemes found so far in the literature all rely on the use of $E - values$ obtained from sequence similarity tools [34, 89, 91, 92, 93], such as BLAST. Very often, a negative $log\ E - value$ between each protein pair is used and some of them incorporate metabolic pathway information from Kyoto Encyclopedia from Genes and Genomes (KEGG) for validating or justifying these scores. The problem with these scoring schemes is that initially there is no single fixed $E - value$ describing where homology ends and non homology begins. This constitutes an impediment to these scoring schemes beyond the fact that they may obviously lead to the singularities caused by the log of zeros.

Thus, these schemes are not adequately equipped to capture all the parameters related to the data under consideration and technology used to derive them, and/or to satisfy numerical computational requirements. In order to overcome these shortcomings, we introduce information-theoretic based measures to score protein-protein relationships in functional interaction networks predicted from sequence data. In these measures, the mutual information of the evolutionary history of protein pairs is corrected by the maximum entropy, measuring how each protein sequence is able to predict the other and thus translating the amount of information shared between proteins into the score of their functional relationships. This allows us to produce a protein-protein functional network from homology data for the MTB genome.

2.1.1 Scoring Scheme For Protein Family and Domain Data

Consider two proteins denoted $\mathcal{P}_i$ and $\mathcal{P}_j$ sharing signatures or entries $\mathcal{S}_1, \ldots, \mathcal{S}_M$. We define the similarity score η_{ij} of proteins $\mathcal{P}_i$ and $\mathcal{P}_j$ as the sum of the minimum number of occurrences of these signatures in proteins $\mathcal{P}_i$ and $\mathcal{P}_j$, *i.e.*,

$$\eta \equiv \eta_{ij} = \sum_{k=1}^{M} \min\{s_{ki}, s_{kj}\} \tag{2.3}$$

where $s_{k\ell}$ is the number of occurrences of signature $\mathcal{S}_k$ in the protein $\mathcal{P}_\ell$.

Broadly speaking, the reliability or confidence score increases with the confidence-level of data, which depends on the data source and decreased with the uncertainty-level of data linked to the dispersion measure σ. As we are dealing with data from experiments containing a certain level of uncertainty, which propagates into the data, it is natural to use the normal distribution, as these data can be summarized in terms of mean and standard deviation. In fact, in this case this distribution constitutes an attractive approximation as it maximizes information entropy in the data. Thus, we set the confidence-level δ of the similarity score η as

$$\delta \equiv \delta(\eta, \sigma) = \phi\left(\frac{\eta^\alpha}{\sigma}\right) \tag{2.4}$$

where the function ϕ is the cumulative probability of the standard Gaussian distribution defined by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{x^2}{2}\right) dx \tag{2.5}$$

and α the control parameter, with $\alpha \geq 1/2$, strengthening the impact of the confidence-level through similarity score η for the data under consideration. And σ the standard deviation of the rectified dataset, estimated from maximum likelihood and given by

$$\sigma = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(x_k - \bar{x}\right)^2} \tag{2.6}$$

where N is the number of signatures found in the rectified dataset, x_k the number of times the signature $\mathcal{S}_k$ was observed, and $\bar{x} = \sum_{k=1}^{N} x_k/N$ the mean or average of the set. The training dataset $\mathcal{D}$ consists of all pairs $(\mathcal{S}_k, x_k)$, where x_k is the number of times the signature S_k was observed. In order to get rid of observations that lie at abnormal distances from the data, referred to as outliers, it is recommended to use the rectified dataset $\mathcal{D}_S$, the subset of the training dataset $\mathcal{D}$ consisting of a data point which falls inside $1.5\,(\mathcal{IQR})$, *i.e.*,

$$\mathcal{D}_S = \left\{ (\mathcal{S}_k, x_k) \in \mathcal{D} : \mathcal{Q}_1 - 1.5\,(\mathcal{IQR}) \leq x_k \leq \mathcal{Q}_3 + 1.5\,(\mathcal{IQR}) \right\}$$

with $\mathcal{Q}_1$ and $\mathcal{Q}_3$ are respectively $1st$ (lower) and $3rd$ (upper) quartile, and $\mathcal{IQR} = \mathcal{Q}_3 - \mathcal{Q}_1$ the interquartile range.

Given the confidence-level δ of the similarity score η defined in equation (2.4), the uncertainty measure related to the outcome η resulting from the data is obtained from the binary entropy function, given by

$$\mathcal{H}_2(\delta) = -\delta \log_2(\delta) - (1-\delta)\log_2(1-\delta) \tag{2.7}$$

In fact, the uncertainty measure function $\mathcal{H}_2(\delta)$ is defined in the interval $[0, 1]$, with $\mathcal{H}_2(0) = 0 = \mathcal{H}_2(1)$ since $\lim_{\epsilon \to 0^+} \epsilon \log_2(\epsilon) = 0$, and also $\lim_{\epsilon \to 1^-} (1-\epsilon)\log_2(1-\epsilon) = 0$.

Finally, we set up the capacity of inferring the functional relationship score between two proteins belonging to the same family or sharing common signatures as

$$\Gamma(\delta) = 1 - \mathcal{H}_2(\delta) \tag{2.8}$$

and the reliability or confidence score of the functional relationship between two proteins by

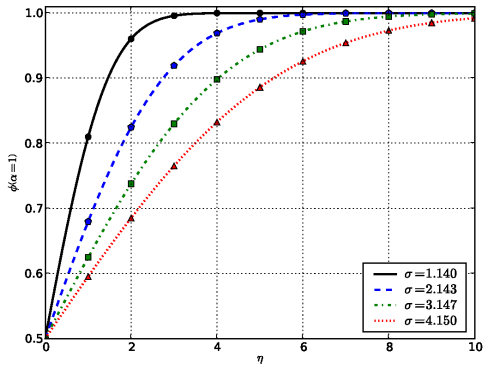
$$\mathcal{R} = \frac{\Gamma(\delta)}{\max_{s} \Gamma(s)} \tag{2.9}$$

Note that for η sufficiently large, δ converges to 1. Therefore, the uncertainty measure $\mathcal{H}_2(\delta)$ converges to 0, leading to the maximum capacity of inferring the functional relationship of 1. This means that the reliability of a functional relationship between two proteins is given by

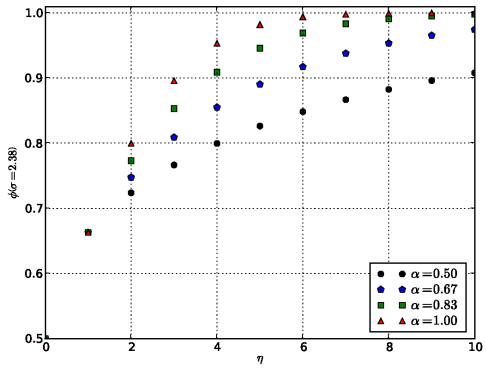
$$\mathcal{R} = \Gamma(\delta)/bit \tag{2.10}$$

To illustrate the dependency of this new measure on the data under consideration and the technology used to produce them, we plot the variation of confidence level δ, uncertainty H_2 and capacity Γ in terms of common domains η between proteins, for different values of α, which keeps track of the technology used to produce data and σ controlling the impact of data under consideration, respectively. These are user-tunable parameters and results are shown in figure 2.1.

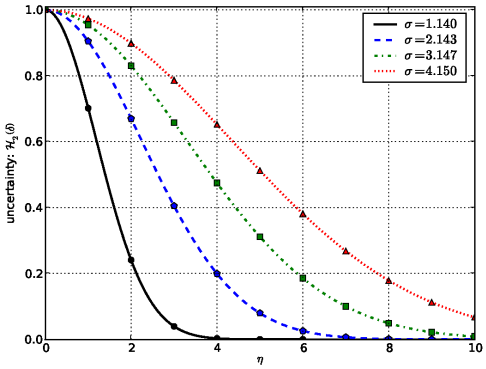
These results show that the confidence level δ increases as the number of common signatures between the two proteins increases, and that for a higher value of α, indicating the efficiency level of the technology used to derive data, the confidence level δ is higher, and so is the reliability or confidence score, due to the fact that in this case the uncertainty component is smaller. Similarly, the impact of data obtained from each technology is taken into account through σ. Interestingly, this confidence score formula accommodates the case where no common pattern is found between two proteins in the training dataset, in which case, the confidence score or reliability of a functional relationship is 0. In addition, this scoring scheme takes into account a false positive assignment of any of the common patterns by narrowing down the confidence score of proteins containing only one common signature, depending on the measure of dispersion σ which can provide a hint on the nature of the data under consideration. Indeed, the measure of dispersion σ impacts on the confidence score in the sense that if data is far away from the average, in which case σ is high, the uncertainty component might be large and significant while calculating the confidence score, thus yielding a lower confidence score. Thus, with knowledge of the data source, the measure of dispersion σ can be penalized by a factor ϵ between 0 and 1, in order to reduce



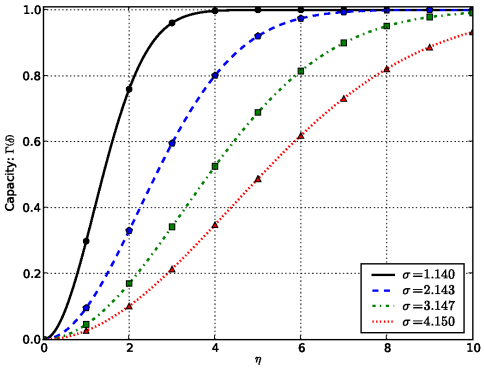
(a) Confidence level variation for $\alpha = 1$.



(b) Confidence level variation for $\sigma = 2.38$.



(c) Variation of uncertainty in terms of σ.



(d) Variation of capacity in terms of σ.

Figure 2.1: *Uncertainty component and reliability variations in terms of tunable parameters α and σ.*

the impact of the uncertainty component.

We compare this new approach to the approach based on protein signature profiling given in equation 2.2, and results are shown in figure 2.2. As the approach based on protein signature profiling may produce several link scores for the same number of shared domains, we have considered the maximum of scores when over-estimating, their minimum when under-estimating and their average. These results indicate that our approach provides a better trade-off between over-estimating and averaging all scores for a given number of shared domains when using the protein signature profiling-based approach.

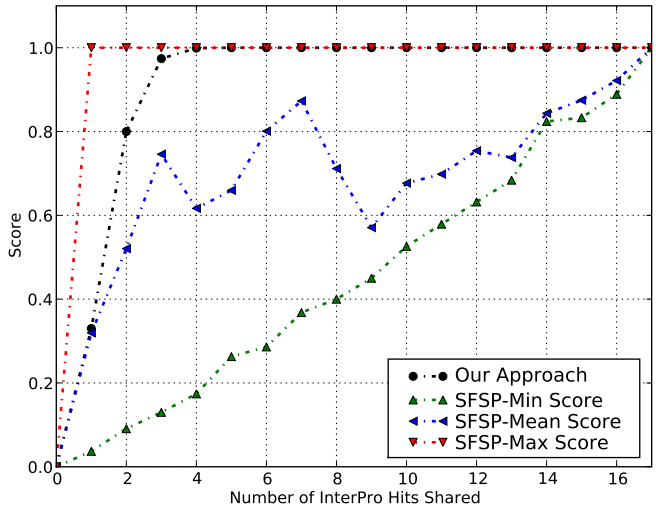


Figure 2.2: *Variation in scores for the Protein Signature Profiling (SFSP) based approach and for our approach.*

2.1.2 Scoring Inferences from Sequence Similarity Data

For a given set of pair-wise homologous sequences, Bastian et al. [94, 95] showed that their biological evolution can be formalized by the evolution of the amount of information they share, which is measured by the mutual information in the sense of Hartley [96, 97], estimating the information they share due to their common origin and parallel evolution under similar selective pressure. Moreover, this mutual information is proportional to the bit score computed with standard methods in sequence comparisons.

Let $S(\mathfrak{s}_1, \mathfrak{s}_2)$ be the bit score alignment of homologous sequences $\mathfrak{s}_1$ and $\mathfrak{s}_2$ set with its standard units, and $\mathcal{M}(\mathfrak{s}_1, \mathfrak{s}_2)$ be mutual information between these two sequences, then we have

$$S(\mathfrak{s}_1, \mathfrak{s}_2) = \vartheta \times \mathcal{M}(\mathfrak{s}_1, \mathfrak{s}_2) \tag{2.11}$$

where ϑ is a constant defining the unity, which depends on the statistical parameter scale K for the search size [98], and derived from a scoring matrix and the amino acid composition of sequence [99]. Therefore, generally $S(\mathfrak{s}_1, \mathfrak{s}_2) \neq S(\mathfrak{s}_2, \mathfrak{s}_1)$ and they are equal only if

they have the same scale for the search size. However, the mutual information $\mathcal{M}(\mathfrak{s}_1, \mathfrak{s}_2)$ between two sequences $\mathfrak{s}_1$ and $\mathfrak{s}_2$ satisfies $\mathcal{M}(\mathfrak{s}_1, \mathfrak{s}_2) = \mathcal{M}(\mathfrak{s}_2, \mathfrak{s}_1)$ and $\mathcal{M}(\mathfrak{s}_1, \mathfrak{s}_2) \geq 0$ [100].

Equation (2.11) shows that the mutual information $\mathcal{M}(\mathfrak{s}_1, \mathfrak{s}_2)$ increases with the bit score $S(\mathfrak{s}_1, \mathfrak{s}_2)$, which measures the average information available per position to distinguish the alignment from chance, calculated as the relative entropy of target and background distributions [101]

$$\mathcal{H}(\mathfrak{s}_1, \mathfrak{s}_2) = \sum_{\imath,\jmath} q_{\imath\jmath} s_{\imath\jmath} = \sum_{\imath,\jmath} q_{\imath\jmath} \log_2 \left(\frac{q_{\imath\jmath}}{p_\imath p_\jmath} \right) \tag{2.12}$$

where $q_{\imath\jmath}$, the "target" residue substitution frequency, is the probability of finding a residue $\imath$ aligned with a residue $\jmath$ after a certain amount of evolution given that they have both evolved from a common ancestor who had a residue k at that position. In addition, $p_\imath$ is the probability of occurrence of a residue $\imath$ in a collection of sequences, *i.e.*, the probability that a residue $\imath$ would align by chance based solely on its frequency in a sequence.

Thus, we define the reliability or confidence score $\mathcal{R}(\mathfrak{s}_1, \mathfrak{s}_2)$ of the functional relationship between two protein sequences $\mathfrak{s}_1$ and $\mathfrak{s}_2$ as the normalized mutual information calculated as

$$\mathcal{R}(\mathfrak{s}_1, \mathfrak{s}_2) = \frac{\mathcal{M}(\mathfrak{s}_1, \mathfrak{s}_2)}{\max\{\mathcal{H}(\mathfrak{s}_1), \mathcal{H}(\mathfrak{s}_2)\}} \tag{2.13}$$

measuring how the protein sequence $\mathfrak{s}_1$ is able to predict the protein sequence $\mathfrak{s}_2$, and where $\mathcal{H}(\mathfrak{s})$ is the relative entropy obtained by aligning protein sequence $\mathfrak{s}$ by itself. Indeed, the increase of mutual information with relative entropy yields bias, and this bias is corrected by dividing the mutual information by the maximum entropy of the sequence pair.

Using equation (2.11), the mutual information $\mathcal{M}(\mathfrak{s}_1, \mathfrak{s}_2)$ can be computed as follows

$$\mathcal{M}(\mathfrak{s}_1, \mathfrak{s}_2) = \frac{S(\mathfrak{s}_1, \mathfrak{s}_2) + S(\mathfrak{s}_2, \mathfrak{s}_1)}{\vartheta + \vartheta'} \tag{2.14}$$

where ϑ and ϑ' are constants defining unity for $S(\mathfrak{s}_1, \mathfrak{s}_2)$ and $S(\mathfrak{s}_2, \mathfrak{s}_1)$ respectively. For a protein sequence $\mathfrak{s}$, $\mathcal{H}(\mathfrak{s}) = \mathcal{M}(\mathfrak{s}, \mathfrak{s})$ obtained using equation (2.14) and given by

$$\mathcal{H}(\mathfrak{s}) = \frac{2 \times S(\mathfrak{s}, \mathfrak{s})}{\vartheta + \vartheta'} \tag{2.15}$$

Finally, $\mathcal{R}(\mathfrak{s}_1, \mathfrak{s}_2)$ is independent of constants defining unity for $S(\mathfrak{s}_1, \mathfrak{s}_2)$ and $S(\mathfrak{s}_2, \mathfrak{s}_1)$, and is calculated as

$$\mathcal{R}(\mathfrak{s}_1, \mathfrak{s}_2) = \frac{S(\mathfrak{s}_1, \mathfrak{s}_2) + S(\mathfrak{s}_2, \mathfrak{s}_1)}{2 \times \max\{S(\mathfrak{s}_1, \mathfrak{s}_1), S(\mathfrak{s}_2, \mathfrak{s}_2)\}} \tag{2.16}$$

It is obvious that this scoring scheme relies only on the two protein sequences for which the confidence score is being computed. Two protein sequences whose mutual information of their evolutionary history embedded in their similarity score is 0, indicates that the two sequences are not similar and so, their confidence score is also 0. Thus, this scoring scheme accommodates the case where no similarity is found between two protein sequences and the error due to the arbitrary growth of the mutual information between two protein pairs is corrected by the maximum entropy induced.

2.1.3 MTB Functional Networks Derived from Sequence Data

We apply these scoring schemes to the whole MTB strain CDC1551 proteome to produce functional links between proteins from sequence data, including pair-wise links from sequence similarity and protein family data derived from the InterPro database. Sequence similarity searches were carried out using BLASTP under a BLOSUM62 matrix based on the premise that if the $E - value$ is less than 0.01, the hit is similar to the query sequence and is likely to be evolutionarily related, and thus is generally suggestive of homology [102]. Sequences in Fasta format and InterPro data for the organism were downloaded from the Integr8 project of the European Bioinformatics Institute (EBI) at `http://www.ebi.ac.uk/integr8`.

The general behaviour of the link confidence scores (score of functional links between protein pairs) induced from sequence datasets has been analyzed and results are shown in table 2.1, in terms of number and frequency of functional links in a given bin $S : x$, where $S : x$ corresponds to the interval $](x-1)/10, x/10]$ of link score values.

		Sequence Similarity		Protein Family and Domain			
Confidence	Bins	Our Approach	STRING scheme	Our Approach	SFSP-Under	SFSP-Aver	SFSP-Over
Low	$S:01$	4321	0	0	33240	0	0
	$S:02$	3001	0	0	4365	0	0
	$S:03$	1206	0	0	814	0	0
	$S:04$	606	44	20915	172	27494	0
Medium	$S:05$	424	263	0	6	6	6
	$S:06$	215	140	0	41	5746	0
	$S:07$	96	99	0	45	1394	0
High	$S:08$	31	57	7847	0	3906	0
	$S:09$	21	58	0	18	155	45
	$S:10$	25	52	9945	6	6	38656
Medium-High Total:		812	669	17792	116	11213	38707
Overall Total :		9946	713	38707	38707	38707	38707

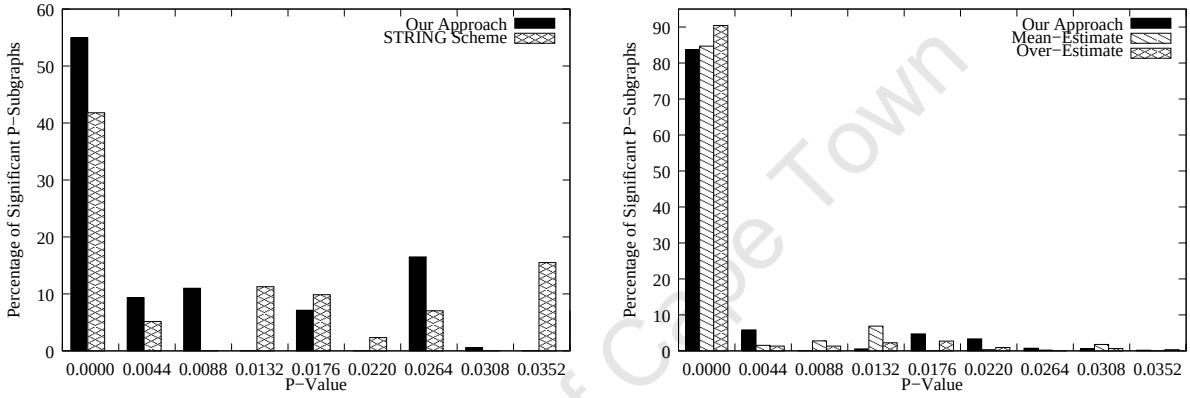
Table 2.1: MTB strain CDC1551 functional links derived from sequence data using our approach, STRING homology scheme for sequence similarity, and using the SFSP approach for protein family and domain sharing. Number of Interactions per Source and Link Score shown separately by bin.

These results indicate that the link confidence scores from protein family data are either low (< 0.4) or high (> 0.7). This is due to the calibration control parameter applied to data from the InterPro database, which is $\alpha = 1$ with penalty parameter $\epsilon = 0.45$, producing either low or high confidence according to the fact that two proteins shared only one domain or more than one domain, respectively. Moreover, in most cases, detection of functional links from sequence similarity matches that of protein family data but at different confidence levels.

2.1.4 Evaluation and Comparison with STRING Scheme

We evaluated the statistical significance and biological relevance of the functional interactions inferred using our scoring approach in terms of functional classification coherence. The functional classes for each MTB protein were extracted from Tuberculist (`http://genolist.pasteur.fr/Tuberculist`), and to determine functional coherence, an interaction between two proteins is said to be significant or correct if these proteins belong to the same functional class. We compared our approach to the STRING homology scoring scheme. The STRING scheme classifies its functional link confidence scores into three different categories, low, medium and high confidence, with corresponding scores less

than 0.4, between 0.4 and 0.7, and greater than 0.7, respectively [81]. The comparison was performed in terms of functional classification accuracy for links with a medium confidence level and upwards (link score greater than 0.4). The number of associations in the MTB functional network are shown separately in table 2.1 for each approach and confidence ranging from low to high. The evaluation was done using a sub-network generated by each protein in the functional network, consisting of functional interactions between a protein under consideration and its direct neighbours, referred to as a P-subgraph. The proteins in the unknown functional class were excluded.



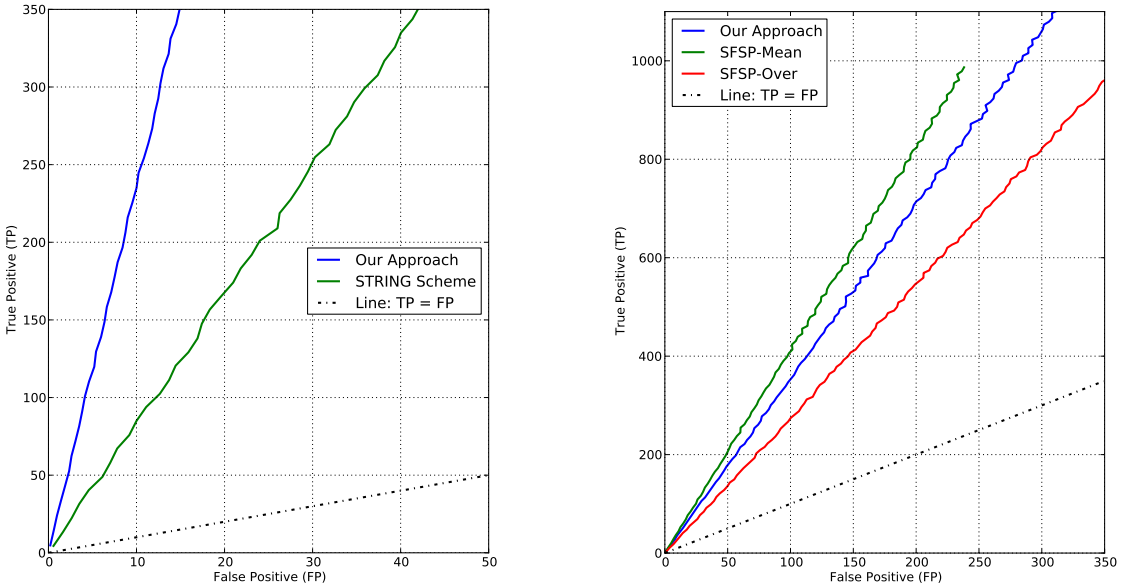
(a) Our Approach and STRING Scheme Evaluation

(b) Our Approach and SFSP Evaluation

Figure 2.3: Significance of Functional Interactions Derived using Our Approach, the STRING scheme and SFSP approach. At each significance level α in these graphs, we counted all relevant predicted associations for the two approaches and computed the percentage. Each α corresponds to the number of associations with p-value β and $\alpha\_ < \beta \leq \alpha$, where $\alpha\_$ is the significance level just before α in the plot.

Statistical Significance of Functional Interactions Derived

To assess functional category coherence of functional interactions derived from a random model, we compute the P-value for each P-subgraph defined as the probability that the P-subgraph under consideration occurs by chance or is comprised of randomly drawn interactions. The hypergeometric distribution, which yields the probability of observing at least ℓ interactions between proteins from a given P-subgraph of size S by chance among I interactions of the same type in the entire functional network considered to be a background



(a) Our Approach and STRING Scheme (b) Our Approach and SFSP

Figure 2.4: Modified ROC curves for functional interactions. Number of incorrect functional interactions (false positives) versus number of correct functional interactions (true positives) in the MTB strain CDC1551 functional networks produced by our approach and the STRING homology network for sequence similarity, and SFSP scheme for protein family and domain.

distribution, is used to model the P-value [91] given by

$$P - value = 1 - \sum_{n=0}^{\ell-1} \frac{\binom{I}{n}\binom{L-I}{S-n}}{\binom{L}{S}} \tag{2.17}$$

where L is the size of the functional network, *i.e.*, the number of functional links in the network, with all the proteins in the unknown class removed.

We assessed functional category coherence of functional interactions derived using our approach and STRING homology data for sequence similarity, as well as those inferred using our scheme for protein family and domain, and those obtained using SFSP-Mean and SFSP-Max estimation. Results displayed in figures 2.3(a) and 2.3(b) show that the functional interactions induced have a very low probability of occurring by chance. Note that this statistical test against a random distribution aims at checking if a given P-subgraph

in the functional network consists of randomly grouped proteins. These figures show that using a significance level of 0.05 as the optimal threshold, more P-subgraphs derived using our approach are statistically significant than those obtained from the STRING homology scoring and provides roughly equal statistically significant percentage of P-subgraphs with SFSP-Mean and SFSP-Max schemes. A total of 205 out of 378, representing 54.2% of P-subgraphs in our network are significant compared to 213 out of 485 representing 43.9% of P-subgraphs for the STRING scoring system for sequence similarity. For SFSP scheme for protein family and domain, A total of 1078 out of 1515 representing 71.2% of P-subgraphs in our network are significant compared to 901 out of 1261 representing 71.5% of P-subgraphs for SFSP-Mean and to 1517 out of 2024 representing 75% for SFSP-Max.

Effectiveness of the Novel Scoring Scheme

To evaluate the classification power of the new scoring scheme, we used the modified Receiver Operator Characteristic (ROC) curve analysis that measures the number of true positive (TP) predictions (number of functional interactions correctly identified) against the number of false positive (FP) (number of functional interactions incorrectly identified) [103], in which case the area under the ROC curve (AUC) is used as a measure of discriminative power. The larger the upper AUC value (the portion between the curve and the line TP = FP), the more powerful the scheme is.

For a given number of P-subgraphs ranging from 5 to 485, we randomly generated 1000 independent samples and compute the average number of correct and incorrect predicted interactions expected to be normally distributed from the central limit theorem. Thus, we perform modified ROC analyses for the two scoring approaches, and results are shown in figure 2.4(a) for sequence similarity. These results indicate that our approach outperforms the STRING scheme, respectively, with an average of 95.9% and 4.1% of functional interactions correctly and incorrectly identified out of 378 P-subgraphs, compared to the STRING scheme, which provides an average of 89.3% and 10.7% of functional interactions correctly and incorrectly identified, respectively, out of 485 P-subgraphs. This shows not only that it is not sufficient to ensure high quality matches [104] by just applying a reasonably strict cut-off score when using the Smith-Waterman algorithm, but also this practice may lead to a poor coverage. Results in figure 2.4(b) indicate that our method performs

comparably to the SFSP-Max and SFSP-Mean schemes, and provides a better trade-off between over-estimating and averaging scores for SFSP schemes in terms of precision and coverage. Our approach provides an average of 79% and 21% of functional interactions correctly and incorrectly, respectively, identified out of 1515 P-subgraphs. SFSP-Mean yields an average of 80.5% and 19.5% of functional interactions correctly and incorrectly identified, respectively, out of 1261 P-subgraphs while SFSP-Max produces an average of 73.3% and 26.7% of functional interactions correctly and incorrectly identified, respectively, out of 2024 P-subgraphs. Apart from the general limitation common to scoring schemes inferred from signature profiling based approaches, SFSP-Max produces a poor precision. This poor performance is due to the fact that when over-estimating it includes all false positives and our approach corrects this, providing an improved precision and coverage.

2.2 Scoring Inferences from Microarray Data

On the basis of the central dogma of biology [105, 106], large scale gene expression mapping finds its origin from the fact that the information on gene expression determines the information on the functional state of an organism. Technologies such as microarrays have been developed to measure the expression level of thousands of genes simultaneously. In fact, microarray technology constitutes one of the most promising tools today to researchers in life science [107], providing an innovative platform for biologists to investigate the dynamic nature of gene dependencies [108]. As a consequence, computational tools are needed in order to extract useful knowledge from such large scale data and to infer relationships between genes from their expression patterns. These vast amounts of data generated at all levels of biological organization can help us to decipher the co-expression network of the genome under consideration. The challenge then lies in tracing the connections and revealing them, *i.e.*, inferring and scoring important gene-gene functional relationships from these data.

Based on the nature of microarray data, usually with a large number of genes and relatively small number of measurements, Principal Component Analysis (PCA) and Partial Least-Squares (PLS) Regression are two methods which can be used to process these data. These two methods have been introduced [84, 109, 110] for microarray data compression,

information extraction and exploration of relationships between genes from their expression patterns. A traditional approach that is very often used, consists of clustering genes using Pearson or Spearman correlation coefficients as distance measures [108] and, in most cases, gene-gene association is characterized by just the correlation coefficient between these genes [111]. However, it is not known how correlated the expression profiles should be in order to infer a relationship between genes, and even knowing the threshold correlation can never provide the evidence of a causal relationship, and more importantly this correlation threshold may only tell us that there exists dependencies between genes without predicting one gene from another. Thus, we decided to make use of a random partial least squares (r-PLS) approach to characterize relationships between genes from microarray data. Each gene is first regressed on all other genes through the r-PLS approach and coefficients are used as a measure (score) of potential relationships between genes, which are later subjected to a statistical test for significance in order to identify outliers. Unlike other approaches in which the complexity of the model is a user-tunable parameter, this approach provides the exact complexity of the model.

Several approaches have been designed for inferring genome-wide co-expression networks of organisms given expression patterns of their genes derived from microarray data, in order to understand normal cell physiology and pathological phenotypes [112]. The description of these approaches can be found in the work published by Markowetz and Spang [113]. Among these approaches, we have encountered partial correlation coefficient methods [114, 115], in which, the partial correlation coefficients are used as a measure of interaction strengths between genes. There are also information-theoretic-based approaches such as ARACNe [112, 116], and Bayesian models such as BANJO [117]. Pihur and colleagues [110] compared these approaches to PLS-based methods, which have been shown to outperform all these approaches and to be a powerful screening tool for discovering interactions among genes from microarray data.

The PLS regression approach has been proven to be a highly efficient statistical regression technique [118, 119], very often the best choice among many statistical methods for linear models with a multicolinearity problem [120], and especially in the presence of noisy data with a limited number of observations [121, 122]. This makes the PLS approach very attractive for the analysis of microarray data which comes with the curse of dimensionality challenges. PLS regression is the dimension reduction technique that finds components

(factors or latent vectors) from predictive or independent, or cause or input variable $\mathcal{X}$, that best predict response or target, or dependent or effect, or output variable $\mathcal{Y}$, by successively extracting factors from both $\mathcal{X}$ and $\mathcal{Y}$ which explain, as much as possible, the covariance between $\mathcal{X}$ and $\mathcal{Y}$ [123]. The PLS regression is quite similar to the PCA approach, the key idea driving these two approaches is the dimension reduction of the original data set by constructing a new set of factors spanning the original data set, without losing essential information. The difference is that the PCA approach decomposes only the predictive variable $\mathcal{X}$ in order to obtain factors which better explain $\mathcal{X}$, *i.e.*, which explain, as much as possible, the variance of $\mathcal{X}$ and no importance is given to how each predictive variable $\mathcal{X}$ may be related to the response variable $\mathcal{Y}$ [124]. So PLS performs one more step allowing cause and effect relationships to be modeled via regression.

Due to the nature of microarray data, in which noise cannot efficiently be controlled, parameterizing the complexity of the model will very often lead to outliers by losing useful information when over- or under-estimating the model. we are using the random partial least squares (r-PLS) approach to produce a MTB co-expression network with exact complexity of the model. Unlike standard PLS in which the number of components or factors is a user-tunable parameter and very often randomly chosen, the r-PLS approach provides a way of optimally finding the number of components needed to represent data under consideration. In this section, we describe briefly the general underlying model of multivariate Partial Least Squares (PLS), and the PLS algorithm in its classical form based on the non-linear iterative partial least squares (NIPALS) algorithm. Thereafter, we set up the statistical test for the significance of the scores (regression coefficients) found. Finally, the r-PLS approach is used to generate the co-expression network of the organism under study from the microarray data extracted from the NCBI GEO [125] and Stanford Microarray [126] databases.

2.2.1 Description of PLS Method and Algorithm

Given respectively data matrices of predictive variables $\mathcal{X}$ and target variables $\mathcal{Y}$ from n observations with k different predictive and m different target variables, the PLS approach is accomplished by finding a linear decomposition of $\mathcal{X}$ and $\mathcal{Y}$ such that

$$\mathcal{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \quad and \quad \mathcal{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \tag{2.18}$$

with $\mathbf{T}^T\mathbf{T} = I = \mathbf{U}^T\mathbf{U}$ and where $\mathbf{T}$ and $\mathbf{U}$, $\mathbf{P}$ and $\mathbf{Q}$, and $\mathbf{E}$ and $\mathbf{F}$ are respectively $n \times r$ scores, $k \times r$ and $m \times r$ loadings, and $n \times k$ and $n \times m$ residuals of $\mathcal{X}$ and $\mathcal{Y}$ as described in figure 2.5. I is the $r \times r$ identity matrix, and T indicates the transpose operator.

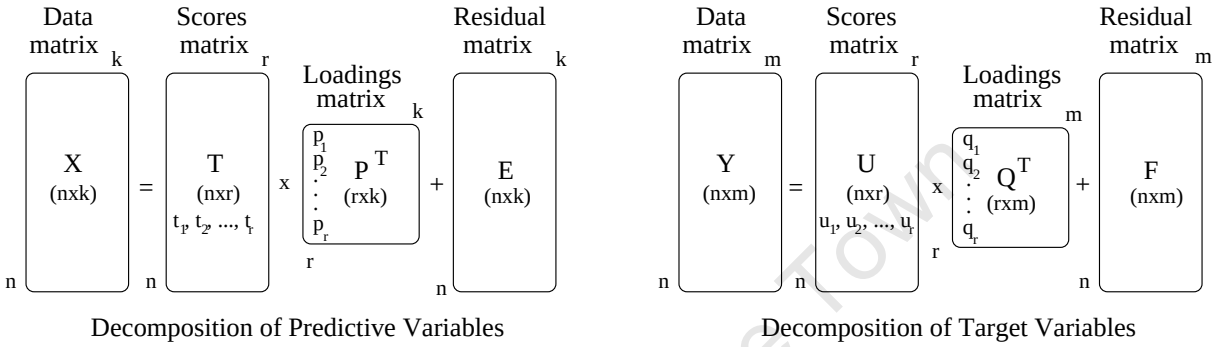


Figure 2.5: *Description of predictive and target variable decompositions.*

PLS Method

The PLS regression is then a stepwise procedure, where at each step the scores t and u are extracted from $\mathcal{X}$ and $\mathcal{Y}$ respectively, reducing the dimension of the predictor and target variables by projecting them to the directions ω and q respectively called input and output weight, that maximize the covariance between input score t and output score u. This is the PLS approach based on the non-linear iterative partial least squares (NIPALS), which consists of finding, at each step, the pair of weights ω and q that solves the following non-linear problem [127].

$$\begin{aligned} &\max\ t^T u \\ &\text{subject to } \omega^T\omega = 1 = q^T q \end{aligned} \tag{2.19}$$

and the scores t and u are given by

$$t = \mathcal{X}\omega \quad and \quad u = \mathcal{Y}q \tag{2.20}$$

So, the decomposition of both predictor and target matrices is done by introducing an input weight ω with the corresponding loading vector p.

NIPALS-PLS Algorithm

Based on the PLS regression principle in finding a pair of weights ω and q, we highlight the PLS algorithm as described by Höskuldsson [127] and used by Rosipal et al. [128] and Abdi [123]. At the beginning the matrices denoted $\mathbf{E} = \mathbf{E}_{(1)}$ and $\mathbf{F} = \mathbf{F}_{(1)}$ are respectively set to the column standardized (centered to have mean 0 and standard deviation 1) data matrices of $\mathcal{X}$ and $\mathcal{Y}$, *i.e.*, $\mathbf{E}_{(1)}$ and $\mathbf{F}_{(1)}$ are column Z-score transformations of $\mathcal{X}$ and $\mathcal{Y}$. This is done in order to reduce data into the same unit scale as shown in figure 2.6.

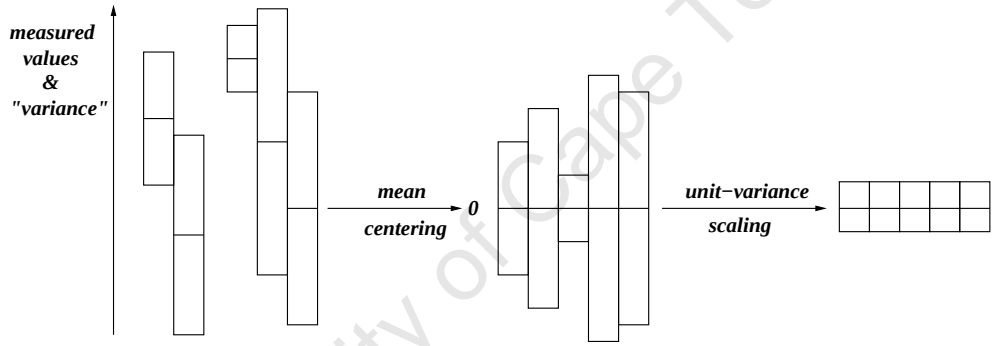


Figure 2.6: *Graphical view of the auto-scaling effect.*

The search for factors or components containing the most information as depicted by figure 2.7 can thus begin.

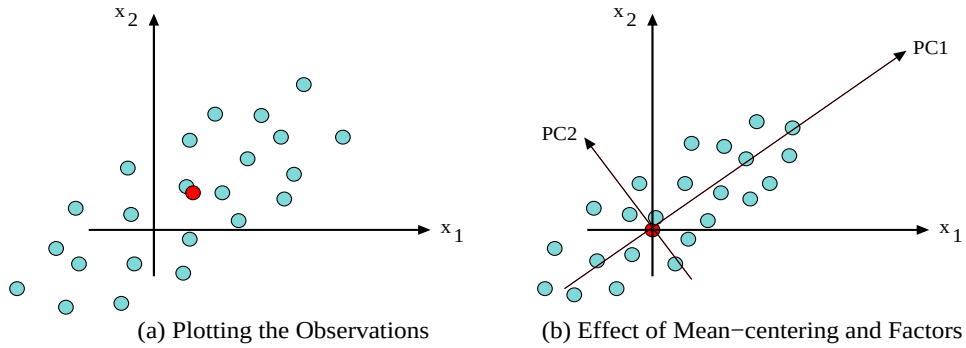


Figure 2.7: *Conceptual representation of factors.*

Figure 2.7(a) shows that x_1 and x_2 do not properly describe the information and (b) emphasizes the fact that the first component PC1 contains the most important information. To obtain these components, we have to find the scores t and u. At each extraction process of these scores, the vector u is initialized with the first column of $\mathbf{F}$ and the following steps are then repeated until convergence:

1. Scale u to be of a unit length, *i.e.*, $\|u\|^2 = u^T u = 1$.
2. $\omega = \mathbf{E}^T u$ and scale ω to be of unit length.
3. $t = \mathbf{E}\omega$ and scale t to be of unit length.
4. $q = \mathbf{F}^T t$ and and scale q to be of unit length.
5. $u = \mathbf{F}q$.

Thereafter, the loading vector p for $\mathcal{X}$ and the value of b used to predict $\mathcal{Y}$ from t are computed as $p = \mathbf{E}^T t$ and $b = \langle t, u \rangle = t^T u$. Assuming that ℓ vectors have been extracted, to move to the next extraction process $\ell + 1$, the deflation process is done by subtracting the effect of t from both $\mathbf{E}$ and $\mathbf{F}$, and repeating the above process with updated matrices $\mathbf{E}$ and $\mathbf{F}$ set to the residual matrices given by

$$\mathbf{E}_{(\ell+1)} = \mathbf{E}_{(\ell)} - t_{(\ell)} p_{(\ell)}^T \quad \textit{and} \quad \mathbf{F}_{(\ell+1)} = \mathbf{F}_{(\ell)} - b_{(\ell)} t_{(\ell)} q_{(\ell)}^T \tag{2.21}$$

where $t_{(\ell)}$, $p_{(\ell)}$, $q_{(\ell)}$ and $b_{(\ell)}$ are parameters obtained at step ℓ. The residual matrix $\mathbf{E}$ becomes null when all the factors have been extracted.

2.2.2 Computational Approach to NIPALS-PLS Algorithm

As observed, the PLS algorithm is built on the properties of the non-linear iterative partial least-squares (NIPALS) algorithm by finding one factor at time. For this purpose, vectors p, q, t, and u are saved as columns in the corresponding matrices at each iteration, and the scalar b is stored as a diagonal element of the diagonal matrix $\mathbf{B}$.

By sequentially substituting the rightmost variable from step 2 of the NIPALS-PLS algorithm, we can see that the input weight ω is proportional to $\mathbf{E}^T\mathbf{F}\mathbf{F}^T\mathbf{E}\omega$ and similarly the output weight q is proportional to $\mathbf{F}^T\mathbf{E}\mathbf{E}^T\mathbf{F}q$. We have

$$\mathbf{E}^T\mathbf{F}\mathbf{F}^T\mathbf{E}\omega = \lambda\omega \quad and \quad \mathbf{F}^T\mathbf{E}\mathbf{E}^T\mathbf{F}q = \mu q \tag{2.22}$$

for certain λ, $\mu \in \mathbb{R}$.

This means that the solution to the PLS problem is to obtain at ω and q, the largest eigenvectors of the matrices $\mathbf{E}^T\mathbf{F}\mathbf{F}^T\mathbf{E}$ and $\mathbf{F}^T\mathbf{E}\mathbf{E}^T\mathbf{F}$ respectively. Thus, since $\mathbf{E}^T\mathbf{F}\mathbf{F}^T\mathbf{E} = \left(\mathbf{E}^T\mathbf{F}\right)\left(\mathbf{E}^T\mathbf{F}\right)^T$ it follows that ω and q are respectively the first right and left singular vector of the matrix

$$\mathcal{S} = \mathbf{E}^T\mathbf{F} \tag{2.23}$$

Thus, computationally ω and q are obtained by using Singular Value Decomposition (SVD) of the matrix $\mathcal{S}$ as first right and left singular vectors respectively. Then, t, p and u are computed as

$$t = \mathbf{E}\omega, \quad p = \mathbf{E}^T t \quad and \quad u = \mathbf{F}q \tag{2.24}$$

and b is the inner product of t and u, *i.e.*, $b = \langle t, u\rangle = t^T u$.

2.2.3 Co-expression Network and Outlier Detection

The PLS method is used to draw meaningful inferences for revealing the regulatory interactions of genes from gene expression data. Interacting genes are more likely to share common genetic control processes and may therefore be functionally related or may at least belong to the same pathway. These interactions are embedded in a co-expression network in which all of these interactions can be indirect through one or more intermediates. A direct link between two genes $\imath$ and $\jmath$ means either one of them is regulating another or the two genes are co-regulated by a third gene. The confidence score or reliability $s_{\imath\jmath}$ of this

link, considered to be the interaction strength between these two genes, is the significant magnitude of the regression coefficient obtained.

PLS Regression Coefficient

With the PLS method as a regression model, the computed factors or latent components are used instead of the original variables. Thus, once $\mathbf{T}$, $\mathbf{P}$, $\mathbf{Q}$ and $\mathbf{U}$ are constructed, the target or response matrix $\mathcal{Y}$ is estimated as the fitted response $\widehat{\mathcal{Y}}$ with the inner relation $\mathbf{U} = \mathbf{TB}$ linking both predictor matrix $\mathcal{X}$ and response matrix $\mathcal{Y}$. This fitted response matrix may be written

$$\widehat{\mathcal{Y}} = \mathbf{TBQ}^T = \mathcal{X}\widehat{\mathcal{B}}_{PLS} \tag{2.25}$$

where the matrix $\widehat{\mathcal{B}}_{PLS}$ of regression coefficients for the model is given by

$$\widehat{\mathcal{B}}_{PLS} = \left(\mathbf{P}^{T+}\right)\mathbf{BQ}^T \tag{2.26}$$

with P^{T+} the Moore-Penrose pseudo-inverse of P^T.

Notice that for starting the NIPALS-PLS algorithm, predictor and response variables are standardized. The non-standardized prediction model is given by

$$\widehat{\mathcal{Y}} = \widehat{\mathbb{B}}_0 + \mathcal{X}\widehat{\mathbb{B}}_{PLS} \tag{2.27}$$

where $\widehat{\mathbb{B}}_0 = \mu_{\mathcal{Y}} + \mathbf{c}_V^T$ and $\widehat{\mathbb{B}}_{PLS} = \left(\sigma_{\mathcal{Y}}\sigma_{-\mathcal{X}}^T\right) \cdot \mathcal{B}_{PLS}$, with $\mu_{\mathcal{Z}}$, $\sigma_{\mathcal{Z}}$ and $\sigma_{-\mathcal{Z}}$ respectively column vectors whose components are mean, standard deviation and standard deviation inverse of each variable in data matrix $\mathcal{Z}$, and the operator "$\cdot$" defines the multiplication in the sense of Hadamard [129]. Additionally, $\mathbf{c}_V = -\sigma_{\mathcal{Y}}^T\left(\Delta_{\mathcal{X}} \cdot \widehat{\mathcal{B}}_{PLS}\right)$ where $\Delta_{\mathcal{X}}$ is the matrix whose rows are equal to $\left(\mu_{\mathcal{X}} \cdot \sigma_{-\mathcal{X}}\right)^T$.

An attractive aspect of the PLS approach is the dimension reduction, the transformation of the numerous original variables into a small number of factors spanning the input

space [120, 130, 131, 132]. This compression is done in such a way that the loss of significant information is minimized. The issue here is to determine the optimal number of factors which provide a well fitting model with significant predictive power. Therefore, strict testing of the predictive significance of each factor when constructing them is necessary, as well as stopping as soon as factors start to become non-significant. To deal with this problem, a cross-validation technique has been proposed and extensively used [123, 133, 134, 135, 136] to determine the correct complexity of a model.

Determination of the Number of Factors

In order to successfully use the PLS method, one has to optimally fix the number of factors to be extracted in the dimension reduction process. The Cross-Validation (CV) technique has been proven to be a practical and reliable way of testing the predictive significance of the PLS model. Specifically, leave-one-out cross-validation has been suggested as a standard technique of determining the optimal number of factors [123, 135, 136].

Assuming that the number of factors is given, referred to as a fixed model, the goodness of fit or the quality of the PLS regression model is measured by the prediction error, referred to as the REsidual Sum of Squares (RESS) and computed as

$$\text{RESS} = \left\| \mathcal{Y} - \widehat{\mathcal{Y}} \right\|_F^2 \tag{2.28}$$

where $\|\mathbf{A}\|_F = \left(\sum_{ij} |\mathbf{A}_{ij}|^2\right)^{1/2} = \left(\text{tr}\left(\mathbf{A}^T\mathbf{A}\right)\right)^{1/2}$ is the matricial Frobenius norm and here tr( ) is the trace function. It follows that the smaller the value of RESS, the better the prediction.

In most cases, the number of factors is not known in advance, this is referred to as a random model. The objective is then to select all possible important variables among numerous and correlated $\mathcal{X}-$variables which sufficiently predict the whole system. For this purpose the leave-one-out or jackknife approach [137, 138], which focuses on the samples that leave out one observation at a time, is widely used to determine the correct complexity of the model, *i.e.*, to determine the optimal number of factors avoiding the risk of "over-fitting" or "under-fitting" the model. The sample obtained by leaving out one observation is called

the jackknife sample, and thus, for n observations, we have exactly n jackknife samples from which n jackknife estimates $\widehat{\mathcal{Y}}_{J_1}, \ldots, \widehat{\mathcal{Y}}_{J_n}$ can be obtained and the jackknife prediction estimate which is the mean of $\widehat{\mathcal{Y}}_{J_1}, \ldots, \widehat{\mathcal{Y}}_{J_n}$ is calculated as

$$\widehat{\mathcal{Y}}_J = \frac{1}{n} \sum_{\ell=1}^{n} \widehat{\mathcal{Y}}_{J_\ell} \tag{2.29}$$

The difference between actual $\mathcal{Y}$ and predicted $\widehat{\mathcal{Y}}_J-$values induced by n jackknife sample models is computed. The cross-validation sum of squares of this difference, calculated from all the jackknife samples, is referred to as the Predictive REsidual Sum of Squares (PRESS), estimating the predictive ability of the model and computed [137] as

$$\text{PRESS} = \left\| \mathcal{Y} - \widehat{\mathcal{Y}}_J \right\|_F^2 \tag{2.30}$$

Thus, when using the cross-validation approach, the ratio $\text{PRESS}_\ell/\text{RESS}_{\ell-1}$ is calculated after each factor and a factor is significant if this ratio is smaller than the threshold set to $(0.95)^2 \approx 0.9$ [123]. $\text{RESS}_{\ell-1}$ denotes the residual sum of squares before the current factor indexed ℓ, with $\text{RESS}_0 = m \times (n-1)$, and the computation of factors continues until a non-significant factor is found.

Significance and Interval Estimation of the PLS Regression Coefficients

Estimating the confidence intervals and significance of the PLS parameters is an active research area [139, 140, 141]. Using a leave-one-out cross validation approach, the variation in the parameters of the various jackknife sample models from data can be used to estimate standard errors, followed by the use of a t-distribution, assuming that $n < 30$ to determine confidence intervals. The jackknife estimate of standard errors of the relationship score $s_{\imath\jmath}$ between genes $\imath$ and $\jmath$ is calculated [137, 138] as

$$\widehat{Se}_{\imath\jmath} = \left[\frac{n-1}{n} \sum_{\ell=1}^{n} \left(\mathcal{B}_{\imath\jmath}^{J_\ell} - \widetilde{\mathcal{B}}_{\imath\jmath} \right)^2 \right]^{\frac{1}{2}} \tag{2.31}$$

where $\mathcal{B}^{J_\ell}$ is the estimate regression coefficient produced from the replicate ℓ-th jackknife sample, $1 \leq \ell \leq n$. And $\widetilde{\mathcal{B}}$ is the jackknife regression coefficient estimate which is the mean of $\mathcal{B}^{J_1}, \ldots, \mathcal{B}^{J_n}$ given by

$$\widetilde{\mathcal{B}} = \frac{1}{n} \sum_{\ell=1}^{n} \mathcal{B}^{J_\ell} \tag{2.32}$$

Thus, at significance level α the confidence limits for $\mathcal{B}_{PLS}$ with $(1-\alpha) \times 100\%$ confidence interval are given by

$$\widetilde{\mathcal{B}} \pm t\left(1 - \frac{\alpha}{2}; n - r\right) \widehat{Se} \tag{2.33}$$

where $\widehat{Se}$ is a matrix whose components are $\widehat{Se}_{\imath\jmath}$, computed in equation (2.31), and referred to as the standard error of $\widetilde{\mathcal{B}}$, and $t\left(1 - \frac{\alpha}{2}; n - r\right)$ is the critical value of the $t-$distribution with probability $1 - \alpha/2$ for $n - r$ degrees of freedom. Note that for the sample size or the number of observation $n \geq 30$, the normal distribution values are used instead of a $t-$distribution in estimation of confidence intervals.

For the significance of the regression coefficient $s_{\imath\jmath}$, the test can be set in the usual fashion. So the test of interest is

$$\mathbf{H}_0 : s_{\imath\jmath} = 0$$
$$\mathbf{H}_a : s_{\imath\jmath} \neq 0$$

Assuming the null hypothesis $\mathbf{H}_0$ is true, we can use the test statistic given by

$$t^* = \frac{\widehat{\mathcal{B}}_{\imath\jmath}}{\widehat{Se}_{\imath\jmath}}$$

where $\widehat{\mathcal{B}}_{\imath\jmath}$ is the estimated regression coefficient and $\widehat{Se}_{\imath\jmath}$ the standard error respectively defined in relations (2.26) and (2.31).

The decision rule is that the coefficient $s_{\imath\jmath}$ is judged significant if $|t^*| > t\left(1 - \alpha/2; n - r\right)$, with α the significance level. Alternatively, one can compute the probability value (p-value), which is the probability of getting a value of the test statistic greater than that observed by chance, assuming the null hypothesis $\mathbf{H}_0$ is true, and $s_{\imath\jmath}$ is judged significant

if the p-value is less than the significance level α. The latter approach has become the standard technique for assessing the significance of predictive parameters in regression analysis, and is incorporated in all available regression tools.

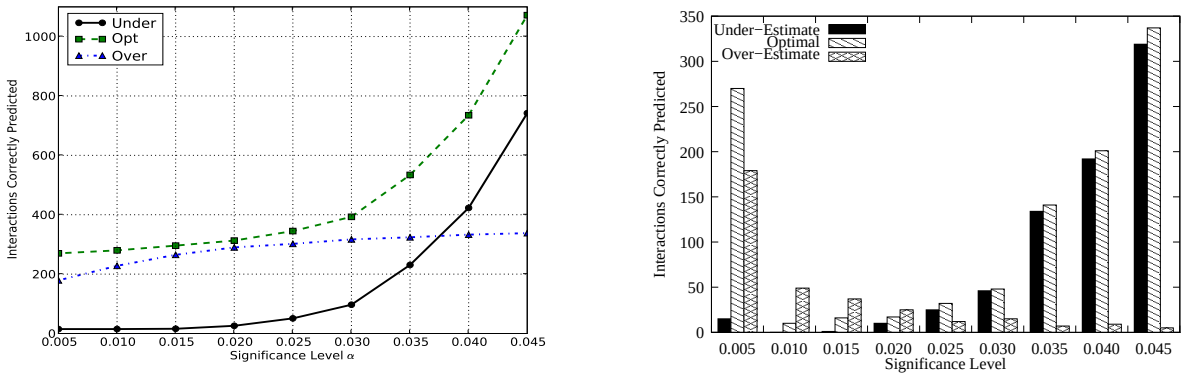
2.2.4 MTB Co-expression Network Derived from Microarray Data

The raw microarray datasets for the organism under study, namely MTB strain CDC1551, were downloaded from the public microarray databases, Stanford Microarray Database (SMD) [126] and NCBI Gene Expression Omnibus (GEO) [125]. We analyzed the standardized log relative intensity ratio, namely the $\log_2(Cy5/Cy3)$ with $Cy5$ (red fluorescent dyes) labelling target and $Cy3$ (green fluorescent dyes) labelling reference samples respectively. To generate MTB co-expression networks using the PLS approach described above, we considered two experiments from the SMD database whose the file Ids are 15569 and 15575, and five experiments from the GEO database whose file Ids are GSM219305, GSM219324, GSM219694, GSM219695 and GSM219696. These files were specifically considered as they provide the highest coverage in terms of total number of common genes when joining the microarray expression data.

Evaluating Functional Interactions Derived

In order to assess the statistical significance of results obtained, we use a significance level $\alpha = 0.05$. The corresponding p-value of gene pair-wise interactions is computed by varying in the parameters of the various leave-one-out sample models from initial datasets. In the co-expression network, an edge linking two genes shows the significance dependency between them if the p-value is less than 0.05.
For biological validation of the model, we use the assumption that a given interaction is correctly predicted or biologically relevant if the two interacting partners belong to the same functional class. For this we used the functional classes assigned to MTB proteins in Tuberculist (`http://genolist.pasteur.fr/Tuberculist`). The number of correctly predicted interactions is then used as a performance measure for comparing the three models, namely under-, over-estimated, and optimal models.

Results are depicted by figure 2.8 in which at each significance level α in these graphs,



(a) Cumulative number of correctly predicted associations.

(b) Number of correctly predicted associations.

Figure 2.8: *Performance analysis of the three models, Under-Optimal-Over Estimated Models.*

we counted all relevant predicted associations for the three models. For the cumulative number of correctly predicted associations, each α in the figure corresponds to the number of associations with p-value $\beta < \alpha$. On the other hand, for the number of correctly predicted associations, each α corresponds to the number of associations with p-value β and $\alpha\_ \leq \beta < \alpha$, where $\alpha\_$ is the significance level just before α in the plot. These results show that in the three models, the number of interactions correctly predicted increases with the significance level α, and the "optimal" model outperforms the under- and over-estimated models.

This is justified by the results shown in table 2.2, where the PLS analysis of MTB microarray data is performed, evaluating on average, variance explained by latent vectors (factors or components). These results show that in most cases, two PLS factors are enough to explain raw MTB data, while one factor was often not enough to extract all useful information from these data, thus missing some of the interactions. In addition, three factors were too many, producing more noise than information needed, thus causing the inefficiency of the under- and over-estimated models, respectively.

Table 2.2: *PLS analysis of MTB raw data.*

# of PLS Factors	% of Explained Var. for Predictive Proteins	Cumulative of Explained Var. for Predictive Proteins	% of Explained Var. for Target Proteins	Cumulative % of Explained Var. for Target Proteins
1	53.68	53.68	92.47	92.47
2	13.17	66.85	6.60	99.06
3	11.52	78.37	0.89	99.96
4	7.36	85.73	0.04	100.00
5	7.04	92.77	0.00	100.00
6	7.23	100.00	0.00	100.00

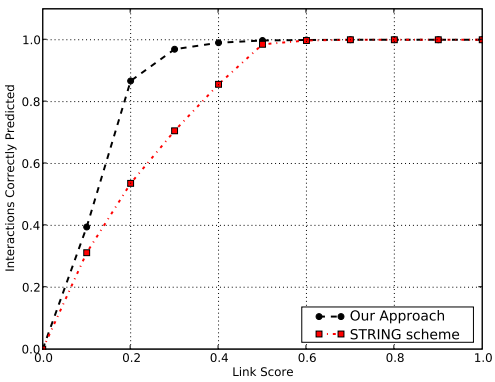
Table 2.3: *Functional interactions in the STRING co-expression network and in our co-expression network.*

	Low Confidence	Medium Confidence	High Confidence
STRING	763	122	0
Our Scheme	6538	225	4

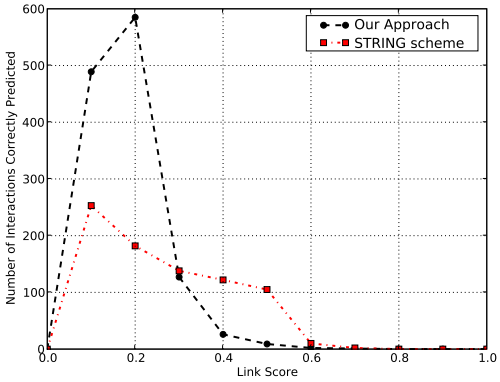
Comparison with the STRING Co-expression Network

The STRING database retrieves its co-expression data from the ArrayProspector web server (`http://www.bork.embl.de/ArrayProspector`), which uses a combination of singular value decomposition and kernel density estimation to compute raw log-odds scores for functional associations between genes with the KEGG map as a reference [84]. The number of functional interactions from the STRING co-expression network and from our approach are shown in table 2.3. The number of biologically significant functional interactions in each interval $]x/10, (x+1)/10]$ of link scores, with $x = 0, \ldots, 9$, is shown in figure 2.9.

Indeed, all the STRING co-expression interactions are contained in our co-expression network and it is not surprising to find that there are no functional interactions with high confidence in the STRING co-expression network. The lower coverage in STRING is due



(a) Cumulative fractions of correctly predicted associations.



(b) Number of correctly predicted associations.

Figure 2.9: *Comparison of functional interactions obtained using our approach to STRING scheme in terms of functional category coherence.*

to the fact that it may not have included all relevant MTB array data available, and it uses the KEGG map as a reference and for the organism under study, for which many of the genes have not yet been classified in KEGG. This also justifies the lack of functional interactions with high confidence in the STRING co-expression network since the STRING scheme assigns a high score to a given functional interaction if the two interacting partners co-occur in at least one KEGG map. For this particular organism, none of the interacting partners have been found in the same KEGG map.

2.3 Summary

We set up scoring systems for identifying and measuring functional relationships between proteins extracted from sequence similarity, protein family and domain and microarray data. These approaches are dynamic, data-driven and technology dependent schemes, producing additional data to STRING for the construction of the MTB functional network. We produced MTB homology and co-expression networks which can contribute to an increase in the confidence and coverage of a unified MTB protein network.

Chapter 3

MTB Proteome Functional Networks

Several biological studies have shown that a protein is a "social animal" [142, 143, 144, 145], *i.e.*, a protein does not achieve its function alone but cooperates with other proteins to perform that function. Therefore, annotating the function of individual proteins in isolation is not sufficient for improving our understanding of their biological processes. Their roles or functions must be integrated with those of other proteins for maintaining the stability of the biological system under unchanging environmental conditions and for the robustness of the system under changing conditions. Many essential cellular processes such as signal transduction, transport, cellular motion and most regulatory mechanisms are thus mediated by protein-protein interactions [146]. These interactions are of various types, but a high level description of biological systems partitions them into two categories, namely physical and functional interactions [92]. Physical interactions refer to physical contact between proteins, and functional interactions or relationships between proteins involve the mechanism through which a particular protein achieves its functions. While "functional interactions" between proteins suggest direct physical contact between them [147], it is actually a broader concept and does not necessarily involve direct physical interactions [146].

In this work, we only refer to functional interactions, including genetic interactions, and those derived from knowledge about co-expression and shared evolutionary history. Proteins interact directly or indirectly through one or more intermediates to carry out their functions in promoting the stability and robustness of the system. These interactions can be modeled as a network, called a protein-protein functional network or interactome. This is a network in which nodes or vertices are proteins and edges or links represent pair-wise

interactions or functional relationships between proteins within an organism. Analytically, protein-protein functional interaction networks are represented as a couple $G(\mathcal{N}, \mathcal{L})$ where $\mathcal{N}$ is the set of proteins (nodes) and $\mathcal{L}$ the set of functional relationships (links), and graphically visualized using an undirected graph layout representing the paths of communication and metabolism of an organism. Even though interaction networks do not directly encode cellular processes nor provide information on dynamics, they do represent a first step towards description of cellular processes, which are ultimately dynamic in nature [148] and they constitute a significant step toward understanding the functional organization of the cell [146]. Therefore, knowledge of protein-protein networks might advance our understanding of biological systems including molecular pathways, elucidate the role of various proteins in complex diseases and how they cooperate to achieve a higher goal in the organism. This could contribute towards improving the control of disease, and thus ultimately enhancing health.

To obtain a protein-protein interaction network, every functional relationship or interaction between proteins should be depicted. These interactions are discovered by various experimental approaches, and often partially complemented with prediction techniques [81]. One of the subjects of heated debate around protein-protein interaction networks is that a network obtained from high-throughput experiments roughly maps the "current" network of interactions occurring inside the cell. Indeed, there are several issues related to high-throughput data including noise, environment and the nature of the approaches used for each experiment [89]. Thus, each specific approach may incorrectly classify interactions, *i.e.*, either failing to detect interactions, referred to as false negatives or wrongly identifying some other interactions, referred to as false positives. The lack of appropriate techniques to address these shortcomings results in biases in the outputs and this is obviously a technology-dependent problem. In order to alleviate the former issue, data integration combining information from multiple interacting data sources into one unified network is deployed, leading to a higher confidence and an increased coverage. For the latter issue, a reliability threshold is applied, thus discarding all functional interactions whose reliability or confidence score is less than the threshold. These techniques are expected to significantly reduce the false negative and positive rate of the network produced, thus yielding a network of high confidence interactions.

In this chapter, we use a computational approach to predict pair-wise functional interac-

tions from several protein-protein interaction datasets in order to produce an integrated MTB proteome network of high reliability interactions using an integrated scoring scheme. A global view of the MTB proteome network is also discussed, analyzing the contribution of each protein-protein interaction dataset to the network produced.

3.1 Data Sources for the Analysis of the MTB Proteome

The recent availability of large amounts of biological data from primary genomic sequences and the exponential growth of high throughput experimental datasets have made possible the analysis of organisms with high confidence, accuracy and precision. These data are currently gathered in the public databases freely accessible via web interfaces, and constitute a rich source of knowledge which has the power to advance our understanding of organisms. With the use of computational approaches, this has opened a new route toward investigating relationships between genes and their products and eventual function prediction of uncharacterized proteins.

In this study, we combine heterogeneous sources of biological data for the analysis of the genome of MTB strain CDC1551. Datasets for the organism under study were downloaded from public biological web sites provided in table 3.1. These data can be categorized into two classes, namely genomic and functional data. Functional data include data from STRING, Gene Ontology (GO), KEGG and MetaCyc pathways, as well as data from high-throughput experiments, such as microarray data. On the other hand, genomic data consists of sequence data including gene and protein sequences, and protein family and domain data.

STRING is collection of known and predicted protein-protein associations derived from high-throughput experimental data, the mining of databases and literature, and from predictions based on genomic analysis for a large number of organisms. As such, it provides a very useful initial overview of the functional partners of a protein, especially for uncharacterized proteins [81, 82]. The Gene Ontology (GO) [87] produces a dynamic and controlled vocabulary for consistently describing the biological context of genes and their products in terms of Molecular Function (MF), Biological Process (BP) and Cellular Component

Table 3.1: *Data resources for the analysis of the MTB proteome.*

Scheme	Description	Types	URL
STRING	Search Tool for Retrieval of Interacting Genes/Proteins	Functional associations inferred from sequence and high throughput data	`http://string.embl.de/`
GO	Gene Ontology	A classification system for annotation of genes and gene products with Molecular Function, Biological Process and Cellular Component	`http://www.geneontology.org`
BioCyc	EcoCyc, MetaCyc and derivatives	A Database architecture for genomes and pathway information	`http://biocyc.org/`
KEGG	Kyoto Encyclopedia of Genes and Genomes	An integrated database of genes and metabolic pathway information for many species	`http://www.genome.jp/kegg /pathway.html`
UniProt	Universal Protein knowledgebase	Centralized resource for protein sequences and functional information	`http://www.uniprot.org/`
InterPro	Integrated documentation resources for protein families, domains and functional sites	An integrated database of motif and domain collections	`http://www.ebi.ac.uk/interpro`
Integr8	Integr8 Project	A resource for genomic and proteomic data	`http://www.ebi.ac.uk/integr8`
SMD	Stanford Microarray Database	A database storing raw and normalized data from microarray experiments	`http://smd.stanford.edu/`
GEO	Gene Expression Omnibus	A database repository of high throughput gene expression data and hybridization arrays, chips, microarrays	`http://www.ncbi.nlm.nih.gov/geo/`

(CC). The systematic analysis of gene and protein biochemical functions is obtained from KEGG [83] and MetaCyc [149], covering most of the known metabolic and regulatory pathways. Microarray data used to construct the co-expression network of the organism under study were downloaded from the public microarray databases Stanford Microarray Database (SMD) [126] and the NCBI Gene Expression Omnibus (GEO) [125].

Genomic data used were obtained from the InterPro and UniProt databases downloaded from the Integr8 project at European Bioinformatics Institute (EBI)(`http://www.ebi.ac.uk/integr8`). The Integr8 project offers an overview of complete genomes and proteomes and has been designed to capture data from different molecular biology databases [150]. Our analysis is performed using the non-redundant complete list of proteins extracted from the UniProt database [151] and protein family data is from InterPro. InterPro [85] integrates together predictive models or signatures representing protein domains, families and functional sites, from multiple source databases, namely, PROSITE, Pfam, PRINTS, ProDom, SMART, TIGRFAMs, PIRSF and SUPERFAMILY, Gene3D, PANTHER [86].

3.2 Functional Interaction Datasets

Analyzing a cell or an organism functioning as an integrated system is becoming more and more important since it allows identification of patterns or properties driving the system. This requires the combination of biological experiments and computational methods in order to determine the complete set of interactions existing between all the proteins in the organism. In this study, we use computational approaches to predict pair-wise protein functional associations depending on the data types, which can be divided into the following three categories.

1. Sequence data by using pair-wise sequence similarity and domain data.
2. Genomic context data, which includes conserved genomic neighbourhood, gene fusion events and phylogenetic distribution patterns.
3. High throughput data providing interactions derived from microarray (co-expression) analysis, text mining, knowledge from pathway databases, and known physical interactions from high throughput experiments.

The functional interactions produced are partitioned into two classes, namely interactions extracted from the STRING database and those derived from sequence and microarray data, using the scoring schemes developed here, and considered to be additional interactions to STRING data. Functional interactions extracted from STRING are comprised of interactions derived from genomic context (genomic conserved neighbour or gene order, gene fusion events and gene co-occurrence or phylogenetic profiles across genomes), text mining, knowledge from pathway databases, and known experimental interactions. Thus, the network is obtained by combining the interactions generated by exploiting 9 different approaches (listed in table 3.2) and each interaction is scored according to the computational approach used to derive it. Understanding the properties of these functional interactions is key to successful mathematical modeling of such a system and developing efficient scoring techniques.

The underlying idea of exploiting the conserved gene neighbourhood or gene order approach to predict functional association between proteins is that proteins whose genes are found to be close to each other across multiple genomes are likely to be involved in the same complex or process or to act in the same functional pathway. Thus, these proteins are expected to interact functionally [152]. This is biologically founded on the fact that proximity between genes is the dominant strategy for identifying an operon in the genome [153]. An operon is a group of genes transcribed in a single messenger RNA (mRNA) and often encodes genes in the same functional pathway [154].

The gene fusion method is another approach that uses the relative positioning of genes in a genome, indicating that if two separate genes in one organism exist as one fused gene in another, then these genes are likely to be functionally related [147]. The reasoning behind this approach is that the fusion of two genes from organism to one another indicates that these genes work together and thus have possibly evolved under common selective pressure and are thus more likely to be functionally linked.

The phylogenetic profile or gene co-occurrence approach models the representation of copies (homologs) of a given gene inherited across organisms during speciation events under strong selective pressure [155]. The presence of a copy in a given organism is represented by 1 and the absence by 0, thus producing the phylogenetic profile or evolutionary path of the gene under consideration across a range of organisms. Proteins with similar phylogenetic

profiles, *i.e.*, present together, are more likely to function together in a functional pathway or protein complex, and they are thus expected to be functionally associated [156].

Text mining is the process of extracting semantically interesting and non-trivial knowledge from unstructured text [157]. Robust data analysis approaches, such as classification and clustering, are used to handle large variations in the extraction of the information from text data. A functional relationship between genes is predicted by representing each gene by a set of abstracts and comparing the sets of abstracts of pair-wise genes to determine their functional relationship based on the frequency of co-occurrence. In STRING, abstracts from PubMed, which contain MEDLINE documents with approximately 11 million biomedical citations covering over 4600 biomedical journals published worldwide [158], are used for systematically searching gene names and aliases [81] on NCBI Entrez (`http://www.ncbi.nlm.niv.gov/entrez`).

The intuition behind interaction data derived using knowledge from pathway databases and known physical interactions is to retrieve interaction data from pathway maps and known physical interaction datasets from high throughput experiments. For this purpose, the STRING database retrieves these interaction data from the KEGG and DIP databases. Note that the Database of Interacting Proteins (DIP: `http://dip.doe-mbi.ucla.edu`) is a database that documents experimentally determined physical protein-protein interactions, providing an integrated set of tools for extracting information about physical protein interaction networks [159]. With respect to KEGG maps, two proteins are considered to be functionally related if they co-occur in at least one KEGG map.
Finally, the basis underlying the other three methods, namely sequence similarity, shared domains and microarray co-expression scoring schemes has been discussed in chapter 2.

3.3 Construction of the MTB Functional Network

A unified scheme has been defined in order to normalize or standardize the protein pair-wise relationship confidence scores from nine different sources taking into account the nature of experiments used to derive them. This scheme has an impact on the final functional similarity between interacting partners, which depends on biological data source. The same type of biological data like sequence or genomic context data may lead to different

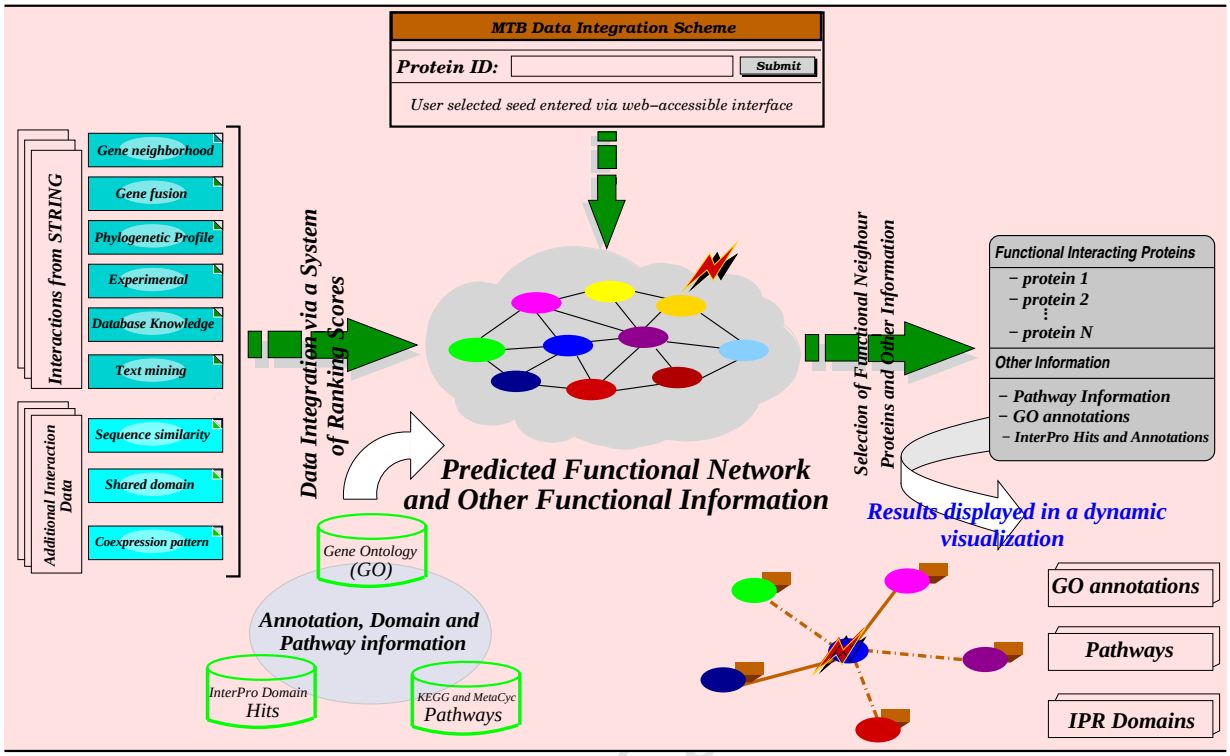


Figure 3.1: *Data accession scheme.*

confidence scores when taking into account the experiment used to derive them. For example, sequence data are more likely to reveal similarity of molecular functions between pair-wise interacting proteins, while genomic context data may indicate sharing of biological process or pathway functional connections rather than exact molecular functions [89]. Nevertheless, as we are expecting our confidence to increase, especially when the functional relationship is predicted from at least two different data types, the combined link confidence score between two proteins i and j for an integrated view of all datasets through a unified network is given by

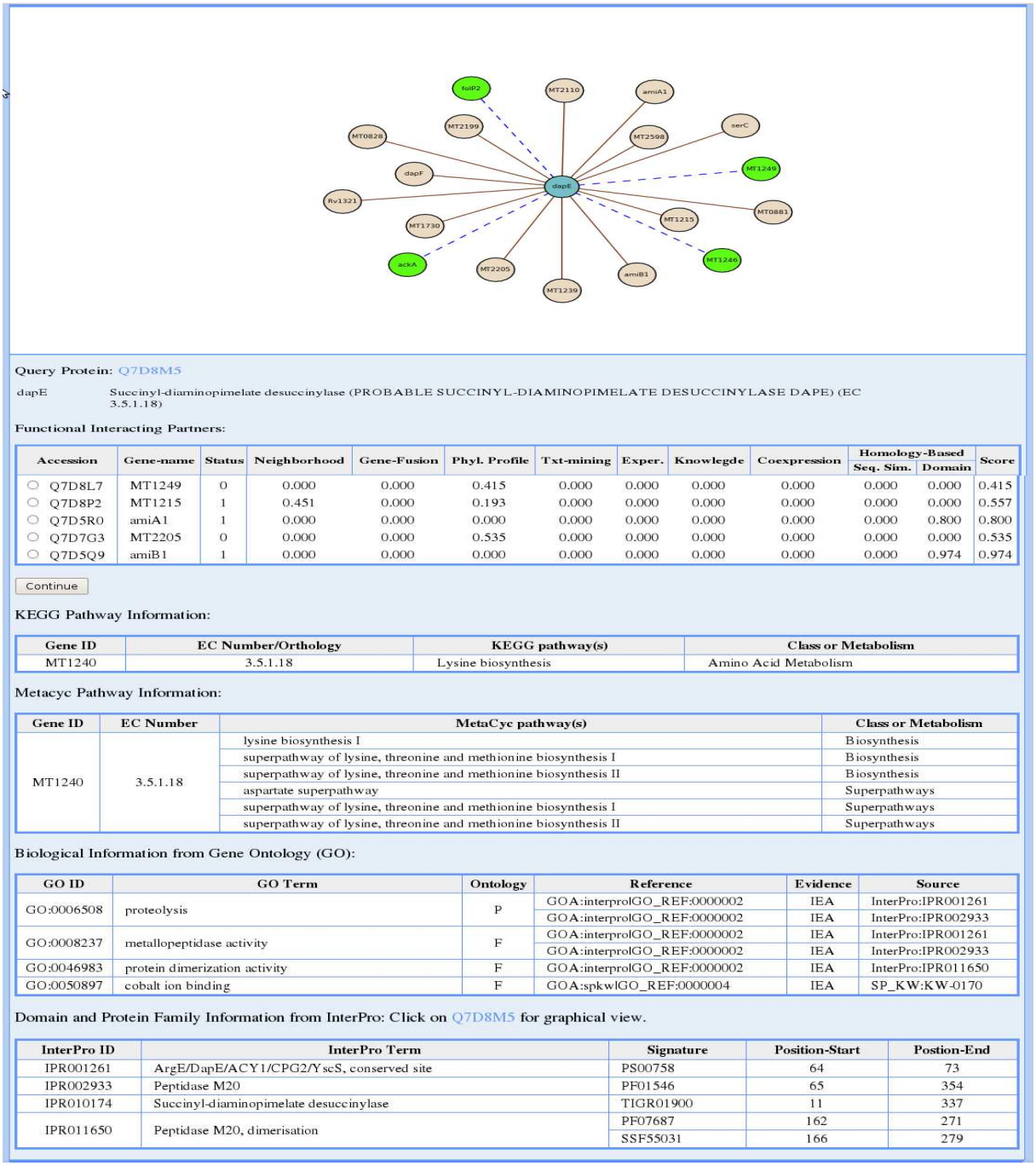
$$\mathcal{S}_{ij} = 1 - \prod_{d=1}^{9} \left(1 - s_{ij}^{d}\right) \tag{3.1}$$

under the assumption of independency, and where s_{ij}^{d} is the confidence score of functional interaction between i and j predicted using the type of data d.

Using this combined score, protein pair-wise functional interactions derived from nine different biological data sources, each with an edge weighing scheme of its own based

on its source, were integrated into a single network and stored in a local MySQL database accessible via a web interface at `http://lab12.cbio.uct.ac.za/tbannotations_v2/`, as described in figure 3.1. Based on this integrated network and a user defined query protein of interest, all proteins in direct relationships with the query protein are identified and displayed together with other functional information related to the query, namely pathway information, annotation from GO and matches to InterPro signatures. The main page of this web interface and an example of the output interface are shown in figures 3.2 and 3.3, respectively.

Figure 3.2: *The main page of the web interface allowing the user access to MTB protein information stored in our MySQL database.*



Query Protein: Q7D8M5

dapE Succinyl-diaminopimelate desuccinylase (PROBABLE SUCCINYL-DIAMINOPIMELATE DESUCCINYLASE DAPE) (EC 3.5.1.18)

Functional Interacting Partners:

Accession	Gene-name	Status	Neighborhood	Gene-Fusion	Phyl. Profile	Txt-mining	Exper.	Knowlegde	Coexpression	Homology-Based		Score
										Seq. Sim.	Domain	
Q7D8L7	MT1249	0	0.000	0.000	0.415	0.000	0.000	0.000	0.000	0.000	0.000	0.415
Q7D8P2	MT1215	1	0.451	0.000	0.193	0.000	0.000	0.000	0.000	0.000	0.000	0.557
Q7D5R0	amiA1	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.800	0.800
Q7D7G3	MT2205	0	0.000	0.000	0.535	0.000	0.000	0.000	0.000	0.000	0.000	0.535
Q7D5Q9	amiB1	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.974	0.974

Continue

KEGG Pathway Information:

Gene ID	EC Number/Orthology	KEGG pathway(s)	Class or Metabolism
MT1240	3.5.1.18	Lysine biosynthesis	Amino Acid Metabolism

Metacyc Pathway Information:

Gene ID	EC Number	MetaCyc pathway(s)	Class or Metabolism
MT1240	3.5.1.18	lysine biosynthesis I	Biosynthesis
		superpathway of lysine, threonine and methionine biosynthesis I	Biosynthesis
		superpathway of lysine, threonine and methionine biosynthesis II	Biosynthesis
		aspartate superpathway	Superpathways
		superpathway of lysine, threonine and methionine biosynthesis I	Superpathways
		superpathway of lysine, threonine and methionine biosynthesis II	Superpathways

Biological Information from Gene Ontology (GO):

GO ID	GO Term	Ontology	Reference	Evidence	Source
GO:0006508	proteolysis	P	GOA:interproIGO_REF:0000002	IEA	InterPro:IPR001261
			GOA:interproIGO_REF:0000002	IEA	InterPro:IPR002933
GO:0008237	metallopeptidase activity	F	GOA:interproIGO_REF:0000002	IEA	InterPro:IPR001261
			GOA:interproIGO_REF:0000002	IEA	InterPro:IPR002933
GO:0046983	protein dimerization activity	F	GOA:interproIGO_REF:0000002	IEA	InterPro:IPR011650
GO:0050897	cobalt ion binding	F	GOA:spkwIGO_REF:0000004	IEA	SP_KW:KW-0170

Domain and Protein Family Information from InterPro: Click on Q7D8M5 for graphical view.

InterPro ID	InterPro Term	Signature	Position-Start	Postion-End
IPR001261	ArgE/DapE/ACY1/CPG2/YscS, conserved site	PS00758	64	73
IPR002933	Peptidase M20	PF01546	65	354
IPR010174	Succinyl-diaminopimelate desuccinylase	TIGR01900	11	337
IPR011650	Peptidase M20, dimerisation	PF07687	162	271
		SSF55031	166	279

Figure 3.3: *An example of the output format for a given protein. The status indicates that a protein is of known functional class, in which case, its status is 1, 0 for those of unknown class or 2 for those of unknown class but predicted to be involved in some functional class.*

Table 3.2: *The number of associations in the MTB functional network, shown separately for each data source and confidence range from low to high.*

Association Evidence by Type	Low Confidence	Medium Confidence	High Confidence
Conserved genomic neighbourhood	1163	6972	4731
Gene fusion events	337	52	99
Phylogenetic Profile	1033	5862	1461
Text mining	1174	722	93
Experimental	220	170	133
Knowledge from database	3	970	2002
Sequence similarity	8524	1345	77
Shared domains	0	20915	17792
Co-expression	6538	225	4
Combined Score	6850	32488	25605

3.4 General View of the MTB Functional Network

We have constructed the MTB functional network from nine biological data sources. The number of functional interactions between proteins derived from these data sources and divided into three different confidence categories, namely low, medium and high confidence is shown in table 3.2.

The final row shows the number of interactions in each confidence range for the final combined score. Note that for a given data source, all interactions whose scores are strictly less than 0.3 (< 0.3) are considered as low confidence, and scores ranging from 0.3 to 0.7 ($0.3 \leq score \leq 0.7$) are classified as medium confidence and scores greater than 0.7 (> 0.7) yield high confidence. Furthermore, the confidence increases when interaction data are integrated into a single network, producing more medium and high confidence links in the last row than when considering only one type of data.

The use of these nine different biological sources is expected to solve the problem of interaction incompleteness. On the other hand, to reduce the impact of bias in functional interactions coming from experimental predictions and computational approaches, we have only considered those ranging from medium to high confidence and for functional interactions with low confidence, only those predicted by at least two different approaches were considered. In total, 5 interactions of low confidence predicted by at least two different approaches have been included in the functional network. We analyzed the network for its

Table 3.3: *General MTB functional network parameters.*

Parameters	Value
Number of Proteins (Nodes)	4136
Number of Functional Interactions (Edges)	58098
Average Degree (in and out)	28
Average Shortest Path Length	3.678
Number of Connected Components	23
% of Nodes in Largest Component	98.7%
Number of Hubs	201

general properties and these network parameters are presented in table 3.3.

The network is comprised of 4136 proteins out of 4195 found in the complete list from UniProt, covering approximately 98.6% of the MTB proteome. Of these, 201 are structural hubs, or "single points of failure", which are able to disconnect the network, thus affecting function, and they are considered to be responsible for the integrity of the system. Due to the presence of these hubs, any pair-wise protein set in a given connected component can communicate through its relative shortest paths. A shortest path between two nodes is the minimum number of hops (edges) required to get to one node from another node. In the MTB functional network, the average path length, obtained by averaging over all shortest paths between all pairs of nodes, is approximately 4 as shown in figure 3.4 representing the probability distribution of the shortest path length.

This reveals that the transmission of biological information from a given protein to others is achieved through only a few steps. Indeed, the average shortest path length value is 3.678, which is approximately of the order of magnitude $\log(|\mathcal{N}|)$ with $|\mathcal{N}| = 4136$. This means that the MTB functional network has a 'small world property' [160, 161] and provides an idea about the network navigability, indicating how fast the information can be spread in the system independently of the number of proteins. This property may also provide the organism with an evolutionary advantage in the sense that the system would be able to efficiently respond to perturbations in the environment and to quickly exhibit a qualitative change of behaviour in response to these perturbations.

We further performed analysis of the degree distribution of the MTB functional network

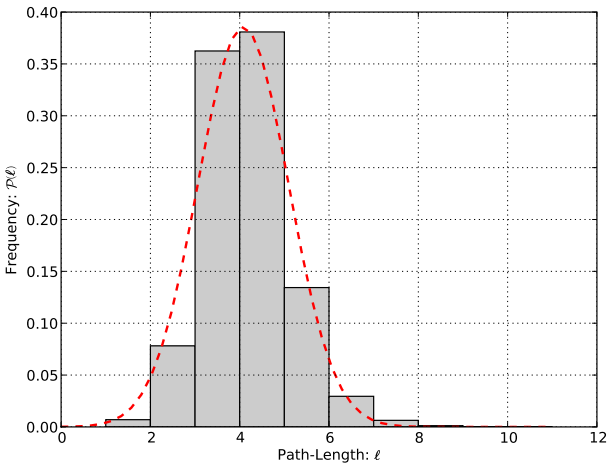


Figure 3.4: *Distribution of shortest path lengths between reachable pair-wise protein functional interactions.*

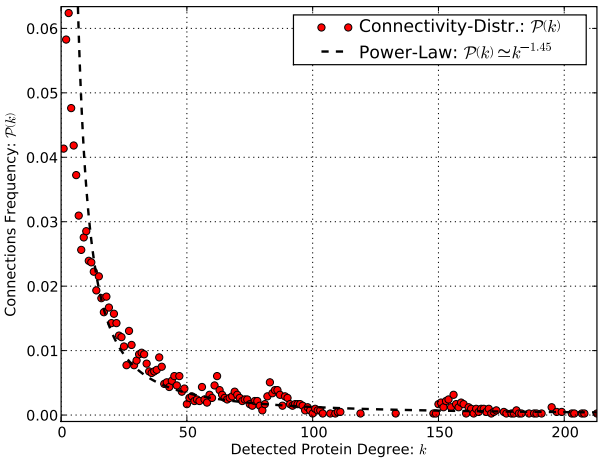


Figure 3.5: *Connectivity distribution of detected k functional links per protein, plotted as a function of frequency* $\mathcal{P}(k)$.

and, as shown in figure 3.5, the functional network exhibits scale-free topology, *i.e.*, the degree distribution of proteins approximates a power law $\mathcal{P}(k) = k^{-\gamma}$, with the degree exponent $\gamma \sim 1.45$. This means that most of the proteins have few interacting partners but some have many partners. The latter are referred to as "high degree nodes", and probably ensure some basic chemical operations such as energy transfer and redox reactions, essential

for the survival of the organism.

3.5 Summary

In this chapter, we integrate functional interactions derived from different biological sources and using computational approaches into a single functional network via the computation of final relationship scores under the independence assumption. This is expected to solve the problem of functional interaction incompleteness of individual data sources. Taking into account the fact that some of these data are unreliable, the functional interactions considered for further analysis are those with a final score ranging from medium to high confidence and those with lower confidence but predicted by at least two different approaches. The global view of the structure of the network produced shows that it satisfies the 'small world' property and exhibits a 'power law' distribution. The network is available for searching via a web interface at `http://lab12.cbio.uct.ac.za/tbannotations_v2/`, which summarizes the neighbourhood and annotation of the query protein.

Chapter 4

Functional Analysis of the MTB Proteome Networks

Proteins perform an astonishing range of biological functions in an organism. These include roles as structural proteins, enzymes and for the transportation of materials within and between cells [78]. Each protein is a gene product that interacts with the cellular environment in some way to promote the cell's growth and function [162]. Therefore, knowledge of protein functions and their biological pathways is crucial for understanding pathogen behaviour and thus for therapeutic study, enabling the development of new drugs.

Despite ever-increasing amounts of biological data including primary data, such as genomic sequences, and functional genomic data from high throughput experiments, there is a deficiency in functional annotation for many newly sequenced proteins. For instance, in most bacterial genomes, as many as 40% of identified proteins are labeled "uncharacterized" or "unknown" or "hypothetical" proteins [163] and specifically, about half of the *Mycobacterium tuberculosis* genome is made up of proteins of unknown functions. This limits the ability to exploit these data, leading to the paradigm of "a world which is data rich yet information poor". Thus, one of the major tasks in the post-genomic era is genome annotation, assigning functions to gene products based mostly on amino acid sequence, in order to capitalize on the knowledge gained through these sequencing efforts. However, experimentally determining the function of these proteins is likely to be difficult for several reasons. These include [164]: (1) possible specific relationship of the function to the native environment in which a particular organism lives, (2) inclusion of many genes in

the genome for securing its survival in a particular environment, which may have no use in the environment created in the laboratory, and (3) it may even, in many cases, be almost impossible to imitate the natural host, with its myriad other micro-organisms, and thereby determine the exact function of gene or gene product by experiment alone. The only effective route toward the elucidation of the function of uncharacterized proteins may be a combination of experimental approaches and predictions through computational analysis.

With the huge amount of data generated over the years, function prediction of identified but yet uncharacterized proteins requires an automated mechanism able to infer functions from related proteins of known functions on the basis of their functional associations or relationships. The framework of our system to achieve this, as described in figure 4.1 follows these steps:

- Generate Functional Interaction Networks enhanced by integrating data from heterogeneous sources (Homology-based, Genomic context and High throughput data).
- Using Gene Ontology (GO) and incorporating pathway information, predict functions of uncharacterized proteins based on the Functional Interaction Networks.

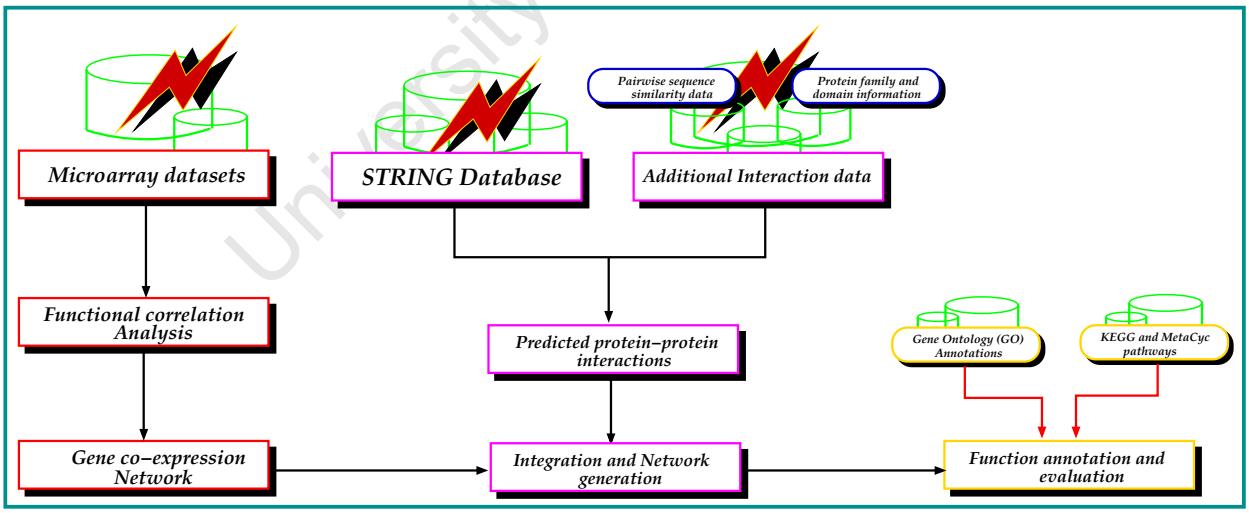


Figure 4.1: *System framework for protein function prediction.*

Predictions along these lines will give a first hint towards functionality that later can be subjected to experimental verification.

The progress made in the use of computational approaches to predict protein function from diverse types of biological data has positively impacted the functional genomics research field. There have been several cases of functional inference in which functions predicted computationally were experimentally validated [165].

Sequence similarity search tools, such as Basic Local Alignment Search Tool (BLAST) [75], have been extensively used for predicting functions of uncharacterized proteins. This approach is referred to as homology-based annotation transfer, providing an easy and straightforward scheme of suggesting possible functions for uncharacterized proteins. The key assumption driving this approach is that two proteins with significantly similar sequences are evolutionary linked and might thus share common functions. However, many factors limit its applicability. For example, no known sequence may be similar to the novel protein sequence in the database, and above all, the most significant database hit may perform a different function due to gene duplication events [142, 166, 167], domain shuffling events (deletions), or single point mutations [168]. As a consequence, on the computational side, the inability of accurately inferring protein functions leads to the propagation of annotation errors in the database [169]. Several approaches that do not rely directly on sequence similarity have also been implemented [170]. These include using information about gene fusions, phylogenetic profiles of proteins families, gene adjacency in genomes and expression patterns. In this work, we are following a data integration approach based on functional networks, which provides the opportunity to exploit all these approaches for characterizing proteins of unknown function.

In the following section we describe the concept of protein function, survey the Gene Ontology (GO), and set up a novel GO semantic similarity metric in order to measure GO term closeness in the hierarchy of GO directed acyclic graph (DAG). Thereafter, we perform functional analysis of the MTB functional network through function prediction, where possible, of uncharacterized proteins in the MTB proteome. We use a prediction approach judged to be the best in terms of quality of prediction and genome coverage. Finally, we provide a new predicted functional classification of the MTB genome strain CDC1551 in terms of GO Molecular Function and Biological Process.

4.1 Protein Function and Gene Ontologies

Mathematically, a set $\mathbf{X}$ is a well-defined collection of objects and a function f from a set $\mathbf{A}$ to a set $\mathbf{B}$ is a rule which associates to each objects (input) $x \in \mathbf{A}$ at most one object (output) $y \in \mathbf{B}$. In this case, y represents the realization of x, called a function of x and denoted $y = f(x)$. Thus, for a function to be well-defined, we need to know the two sets $\mathbf{A}$ and $\mathbf{B}$ and the rule of associations of objects or realizations of all objects of $\mathbf{A}$. In addition, if the sets $\mathbf{A}$ and $\mathbf{B}$ are well described, a function f is completely determined by knowing just the realizations of objects. Similarly, assuming the context and the scope of interest are known, protein function is a concept used to describe all type of realizations or activities to which the protein contributes.

Indeed, even from a biological point of view, the activity to which a protein contributes takes place within an organism and has consequences ranging from the cellular level to the whole organism. This means that the "protein function" concept has various aspects. Firstly, the location where a protein is performing its function must be known and then its contribution has to be categorized. For instance, a protein may be considered to be an enzyme in a given biochemical reaction inside the cell and this is referred to as the biochemical or molecular function of a protein. A protein may contribute to complex physiological reactions within *metabolic* pathways as a component of a complex or via interactions with other proteins in a cell, guaranteeing an efficient functioning of the organism. This is referred to as its cellular function. Furthermore, as a component of an integrated physiological system, a change in protein due to the response of the system to perturbations or environmental stimuli determines the phenotypic output of the organism; this is referred to as the phenotypic function [171]. This demonstrates the subjectivity and ambiguity of the concept "protein function" without describing the context and the scope of interest. Therefore, protein function assignment requires the characterization of protein contributions using well-defined and structured vocabularies specifying the aspect and the context surrounding these contributions.

4.1.1 Description of Protein Function

As mentioned previously, for good characterization of protein function we need to describe its attributes in a systematic manner using a standardized syntax and semantics in a format that is human readable and understandable, as well as being interpretable computationally. The terms used for describing a function should have definitions and be placed within a structure of relationships [172]. Therefore, an ontology is required in order to represent annotations of known genes and proteins, and to predict functional annotations of those which are identified but so far uncharacterized. An ontology is an explicit specification of concepts that includes a set of objects, their properties and their values along with describable relationships between them. This is reflected in a representational vocabulary for a specific domain, containing definitions of classes, relations, functions and other objects [173, 174, 175, 176, 177]. An ontology may be simplified in a hierarchical classification showing the type subsumption relations between concepts in the field of discourse, and visualized as an abstract graph with nodes and labeled arcs representing the objects and relations.

By capturing the knowledge about a domain in shareable and computationally accessible form, ontologies can provide defined, accessible and computable semantics about the domain knowledge they describe [172]. Thus in biology, ontologies are expected to produce an efficient and standardized functional scheme for describing genes and gene products. The earliest controlled vocabulary and well-defined protein function relationships scheme arose from the biochemistry field in the form of the Enzyme Commission (EC) classification [178]. An EC term is used to classify enzymes in biochemical reactions from a unique identifier with a four-level hierarchy of the form "EC -.-.-.-", starting from the more general in the first position, representing a class of enzymes responsible for the catalysis of metabolic reactions, to the more specific in the fourth position specifying the precise biochemical reaction in which a particular enzyme is involved. Unfortunately, this scheme is essentially limited to the classification of protein enzymatic function in biochemical reactions. Thus, several other functional schemes have been suggested [179, 180] for a wider class of genes and gene products with the common objective of setting up an ontology for gene annotations. Structured and controlled vocabularies provide the ability to explore functional annotations of genes and their products in an automated fashion. Although these ontolo-

gies are currently more generalized and more widely applicable due to the unification of biology and the information about genes and proteins shared by different organisms, most of them were initially designed for specific organisms in order to exploit the biological properties of their genomes. For example, EcoCyc [181, 182, 183] was originally set up to categorize gene products of *Escherichia coli K*12 (*E. coli*).

Generally, a more widely applicable ontology should be designed to cover a wide range of organisms, ensuring the integration of biological phenomena occurring in a wide variety of biological systems. In addition, it must be dynamic in its nature in order to enable the design to incorporate new knowledge of gene and protein roles over time. One of the biggest accomplishments in this area is the creation of the Gene Ontology (GO) [87], which currently serves as the dominant and most popular functional classification scheme [91, 184] for annotation and functional representation of genes and their products. In this work, we make use of GO terms to predict, where possible, functions of a large number of proteins of unknown function in *Mycobacterium tuberculosis*.

4.1.2 Gene Ontology (GO)

The necessity for organizing and unifying biology and information about genes and proteins shared by different organisms has led to the construction of the Gene Ontology (GO) [87]. At its outset, GO aims at producing a dynamic, structured and controlled vocabulary describing the role of genes and their products in any organism, thus allowing humans and computers to resolve language ambiguity.

GO provides three key biological aspects about genes and their products in a living cell, namely, the complete description of the tasks that are carried out by individual proteins, their broad biological goals, and the subcellular components, or locations where the activities are taking place. Thus, GO consists of three distinct classification schemes, molecular function (MF), biological process (BP) and cellular component (CC) [185], each engineered as a Directed Acyclic Graph (DAG), allowing a term (node) to have more than one single parent in order to characterize those involved in several molecular functions, biological processes, and subcellular locations. Traditionally, there are two types of relationships between a parent and a child [186]. The "is_a" relation meaning that a child is a sub-

Table 4.1: *GO evidence codes.*

Class	Category	Evidence
Manually-assigned	Experimemtal	EXP: Inferred from Experiment
		IDA: Inferred from Direct Assay
		IPI: Inferred from Physical Interaction
		IMP: Inferred from Mutant Phenotype
		IGI: Inferred from Genetic Interaction
		IEP: Inferred from Expression Pattern
	Computational Analysis	ISS: Inferred from Sequence or structural Similarity
		ISO: Inferred from Sequence Orthology
		ISA: Inferred from Sequence
		ISM: Inferred from Sequence Model
		IGC: Inferred from Genomic Context
		RCA: Inferred from Reviewed Computational Analysis
	Author Statements	TAS: Traceable Author Statement
		NAS: Non-traceable Author Statement
	Curatorial Statements	IC: Inferred by Curator
		ND: No biological Data available
Automatically-assigned		IEA: Inferred from Electronic Annotation
Obsolete		NR: Not Recorded

class or an instance of the parent, and the "part_of" relation indicating the child is a component of a parent. Thus, each edge in a GO DAG represents either an "is_a" or a "part_of" association between terms, and each term (node) has an unique identifier of the form "GO:xxxxxxx", with "xxxxxxx" a zeros padded integer of seven digits [187]. Another relationship has emerged, namely "regulates", which includes "positively_regulates" and "negatively_regulates", and provides for relationships between regulatory terms and their regulated parents [186, 188]. As we are only interested in the GO DAG topology, we only refer to the relations 'is_a' and 'part_of' here, and these are treated equally. Unless specified explicitly, in the rest of this work $\mathcal{N}_{GO}$ and $\mathcal{L}_{GO}$ will respectively express the set of GO terms and links, and $[x, y] \in \mathcal{N}_{GO}$ indicates that the level of term x is lower than that of y. Considering the daily updates of the GO database, we need to mention that the GO data used here were downloaded from the GO database on the 20^{th} of October 2009.

The GO has been widely used and deployed in several protein function prediction analyses

in genomics and proteomics. This growth of popularity is mainly owed to the fundamental organization principles and functional aspects of its conception displayed by its wide coverage and biological relevance. Specific tools, such as the AmiGO browser [189, 190], have been developed for making GO easy to use, and have significantly contributed to the large expansion of GO in the experimental and computational biology fields. Nowadays, GO is the most widely adopted ontology by the life science community [191], and this superiority of GO has been proven by successes resulting from its use in protein function prediction. The GO annotation project arose in order to provide high-quality annotations to gene products, and is applied in the UniProt knowledgebase (UniProtKB) and International Protein Index (IPI) [192, 193, 194, 195]. It also provides a central dataset for annotations in other major multi-species databases, such as Ensembl and NCBI [196, 197]. The reliability or confidence score of the annotation is embodied through evidence codes indicating how the annotation was assigned. The GO Consortium divides these evidence codes into three classes [198], manually-assigned, automatically-assigned and obsolete evidence codes; and manually-assigned evidence codes, in turn, fall into four categories: experimental, computational analysis, author statements, and curatorial statements. Table 4.1 presents a complete list of these evidence codes. Note that among these evidence codes, only IEA (Inferred from Electronic Annotation) is not assigned by a curator and is therefore of the lowest quality.

4.2 Structuring GO for Protein Function Prediction

Considering the wide use of GO, the issues related to its design and usage have been qualified as critical points [199] to be taken seriously for effectively deploying GO in genome annotation analysis. One of the issues is associated with the depth of GO, which often reflects the vagaries in different levels of biological knowledge, rather than anything intrinsic about the terms [172]. Consequently, two genes or proteins may be functionally identical or similar, but technically annotated and labelled with different GO Ids. If only exact matches to GO terms are used in function prediction then these cases are lost. Although several approaches have been designed to assess the similarity and correlation between genes [200, 201, 202, 203, 204, 205] using their sequences or gene expression patterns from high throughput biology technologies, there is still a lack of an effective approach for

discovering functional similarities of genes based on their GO annotations derived from heterogeneous data sources. Indeed, an effective approach should be able to consider the issue related to the depth of the GO DAG (for example where a path goes singly down the hierarchy without branching off) and provide a clear relation of how similar a parent and child are using only the GO DAG topology. This should apply to gene or protein GO annotations derived from different sources, and be independent of the size of the GO DAG, as GO is expanding and increasing in size.

4.2.1 Existing GO Semantic Similarity Measures

Several GO-term similarity measures have been proposed for characterizing similar terms, each having its own strengths and weaknesses. These similarity measures are partitioned into edge- and node-based approaches according to Pesquita et al. [191]. While edge-based similarity measures are based mainly on counting the number of edges in the graph to get the path between two terms, they suffer from the fact that nodes and edges are uniformly distributed, and edges at the same level correspond to the same semantic distance between terms. The node-based approaches use a concept of information content, also called semantic value, to compare the properties of the terms themselves and relations to their ancestors or descendants, and these measures are referred to as IC-based (Information Content-based) approaches [206]. We are interested in IC-based approaches and, unlike the graph-based or hybrid approach introduced by Wang [207], which is based on the intrinsic structure of the GO DAG, *i.e*, only uses the GO DAG topology to compute the semantic similarity, other measures do not consider only the topology of the GO DAG. Most of them are adapted from Resnik [208] or Lin's [209] methods, in which the information content (or semantic value) of a term conveying its biological description and specificity is based on the annotation statistics related to the terms [172, 210], and thus they have a natural singularity problem caused by orphan terms. Here these approaches are referred to as Resnik-related approaches, whose information content or semantic value of a given term z in the ontology, is computed as

$$IC_{RL}(z) = -\ln(p(z)) \tag{4.1}$$

where $p(z)$ is the frequency of occurrence of the term z in the genome under consideration. Resnik's semantic similarity between two terms x and y is given by

$$\mathcal{S}_R(x,y) = \max_{z \in \mathcal{A}(x,y)} \left(-\ln\left(p(z)\right)\right) \tag{4.2}$$

where $\mathcal{A}(x,y)$ is the set of common ancestors to terms x and y, and that of Lin is calculated as

$$\mathcal{S}_L(x,y) = \max_{z \in \mathcal{A}(x,y)} \left(\frac{2 \times \ln\left(p(z)\right)}{\ln\left(p(x)\right) + \ln\left(p(y)\right)}\right) \tag{4.3}$$

Therefore, the more often the term is used for annotation, the lower its semantic value, and as pointed out by Wang, this may lead to different semantic values of the GO terms for heterogeneous data, whereas each biological term in the ontology is expected to have a fixed semantic value when used in genome annotation. The semantic value is defined as the biological content of a given term, and this may be a serious issue. This is especially a problem due to the hierarchical structure of the GO DAG if the information will be used to predict functions of uncharacterized proteins in the genome, in the sense that one source can annotate a given protein with a term at a low level and another source with a term at a higher level in the hierarchy. Furthermore, the description and the specificity of a given term in GO essentially depends on its GO annotation specification, translated by its position in the GO DAG structure or topology.

To overcome these limitations, Wang introduced a topology-based semantic similarity measure in which the semantic value of a given term z is given by

$$IC_W(z) = \sum_{t \in T_z} S_z(t) \tag{4.4}$$

where T_z denotes the set of ancestors of the term z including z, and $S_z(t)$ is calculated as follows:

$$S_z(t) = \begin{cases} 1 & \text{if } t = z \\ \max\{\omega_e * S_z(t') : t' \in \mathcal{C}_h(t)\} & \text{otherwise} \end{cases} \tag{4.5}$$

with $\mathcal{C}_h(t)$ the set of children of the term t, and ω_e the semantic contribution factor for 'is_a' and 'part_of' relations set to 0.8 and 0.6, respectively. The semantic similarity of the two GO terms is given by

$$\mathcal{S}_W(x,y) = \frac{\sum_{t \in T_x \cap T_y} (S_x(t) + S_y(t))}{IC_W(x) + IC_W(y)} \tag{4.6}$$

It has been shown that the Wang et al. approach performs better than Resnik's approach in clustering gene pairs according to their semantic similarity [207, 206].

However, a general limitation common to all these semantic similarity measures is that none of them fully address the issue related to the depth of the GO DAG as stated previously, *i.e*, the depth sometimes reflects vagaries in different levels of knowledge. An example is where the structure is just growing deeper in one path without spreading sideways, in which case terms going down the DAG should have the same semantic value or be topologically identical. Here we are introducing a new semantic similarity measure of GO terms, referred to as topological information of GO terms and based only on the GO DAG topology, to determine the functional closeness of genes and their products based on the semantic similarity of GO terms used to annotate them. This measure incorporates position characteristic parameters of GO terms to provide an unequivocal difference between more general terms at the lower level, or closer to the root, and more specific terms at the higher level, or further from root node. Furthermore, this new measure is a hybrid node- and edge-based approach, overcoming not only the issue related to the GO DAG depth, as stated above, but also the issues related to the dependence on the annotation statistics of node-based approaches and those related to edge-based approaches in which nodes and edges are uniformly distributed.

4.2.2 GO-Universal Semantic Similarity Metric

To overcome the issue of depth of the GO DAG sometimes just reflecting the vagaries of biological knowledge, we are introducing a topological identity or synonym term measure referred to as its topological information. This measure is used to define a novel GO-term semantic similarity measure in order to ensure effective exploitation of the large amounts of biological knowledge that GO offers. This, in turn, provides a measurement of functional similarity between proteins on the basis of their annotations from heterogeneous data using semantic similarities of their GO terms, and thus provides an improved tool for genome

annotation.

Topological Information of a GO Term

Translating the biological content of a given GO term into a numerical value, called the semantic value or topological information, on the basis of its location in the GO DAG, requires complete knowledge of its immediate parents' topological position characteristics. This leads to a recursive formula for measuring topological information of a given GO term, in which the child is eventually expected to be more specific than its parents. The more children a term has, the more specific its children are compared to that term, and the greater the biological difference. Also, the more parents a term has, the greater the biological difference of this term to each of its parent terms. The three separate ontologies, namely, Molecular Function (MF), Biological Process (BP) and Cellular Component (CC) with GO Ids GO:0003674, GO:0005575 and GO:0008150 respectively, are roots for the complete ontology, located at level 0, considered to be the reference level, and are assumed to be biologically meaningless.

Definition 4.1. *The topological information* $\tau(z)$ *of a given term* $z \in \mathcal{N}_{GO}$ *is computed as*

$$\tau(z) = -\ln(\mu(z)) \tag{4.7}$$

★

where $\mu(z)$ is the topological position characteristic of z, recursively obtained using its parents gathered in the set $\mathscr{P}_z = \{x : (x, z) \in \mathcal{L}_{GO}\}$, and given by

$$\mu(z) = \begin{cases} 1 & \text{if } z \text{ is a root} \\ \prod_{x \in \mathscr{P}_z} \dfrac{\mu(x)}{\eta_x} & \text{otherwise} \end{cases} \tag{4.8}$$

with η_x the number of children of term x as parent, and calculated as follows

$$\eta_x = \sum_{y:(x,y) \in \mathcal{L}_{GO}} \delta_x(y)$$

and δ_x is an $x-$children indicator function, *i.e.*,

$$\delta_x(y) = \begin{cases} 1 & \text{if } y \text{ is a child of } x, \textit{ i.e.}, \ (x,y) \in \mathcal{L}_{GO} \\ 0 & \text{otherwise} \end{cases}$$

The topological position is thus a function $\mu : \mathcal{N}_{GO} \longrightarrow [0,1]$, such that for any term $t \in \mathcal{N}_{GO}$, $\mu(t)$ defines a reachability measure of an instance of term t. And obviously, μ is monotonically increasing as one moves up the three, that is if t_1 is_a t_2 then $\mu(t_1) \leq \mu(t_2)$. For the top node or root, the reachability measure is 1.

Definition 4.2. *Let $[x,y] \in \mathcal{N}_{GO}$, x and y are topologically identical or synonym terms, and denoted $x \overset{GO}{=} y$, if the following properties are satisfied.*

- $\tau(x) = \tau(y)$ *or* $\mu(x) = \mu(y)$
- *There exists one path p_{xy} from x to y.* ★

So, two GO terms are equal if and only if they are either the same or topologically identical terms. Suppose that there exists a path p_{xy} from term x to term y, x is a more general term compared to y or y is more specific compared to x, and denoted $x \overset{\text{GO}}{<} y$ if $\tau(x) < \tau(y)$ or $\mu(y) < \mu(x)$.

Furthermore, the topological position μ provides a new way of assessing the intrinsic closeness of GO terms. Two terms in the GO DAG may share multiple ancestors as a GO term can have several parents through multiple paths. So, assuming that a given term x is its own ancestor, we define the topological position $\mu_s(x,y)$ of x and y as that of their common ancestor with the smallest topological position characteristic, *i.e.*,

$$\mu_s(x,y) = \min\{\mu(t) : t \in \mathcal{A}(x,y)\} \tag{4.9}$$

where $\mathcal{A}(x,y)$ is the set of ancestral terms shared by both terms x and y, and containing x and y. Finally, the semantic similarity score of the two GO terms is given by

$$\mathcal{S}_{GO}(x,y) = \frac{\tau(x,y)}{\max\{\tau(x), \tau(y)\}} \tag{4.10}$$

with $\tau(x,y) = -\ln \mu_s(x,y)$ the topological information shared by the two concepts x and y.

Clearly, the values of this semantic similarity measure range between 0 and 1, *i.e.*, $0 \leq \mathcal{S}_{GO}(x,y) \leq 1$, and for any GO-terms x, y, and z in the GO-DAG, $\mathcal{S}_{GO}$ satisfies the following property:

$$\mathcal{S}_{GO}(x,z) + \mathcal{S}_{GO}(z,y) \leq 1 + \mathcal{S}_{GO}(x,y) \tag{4.11}$$

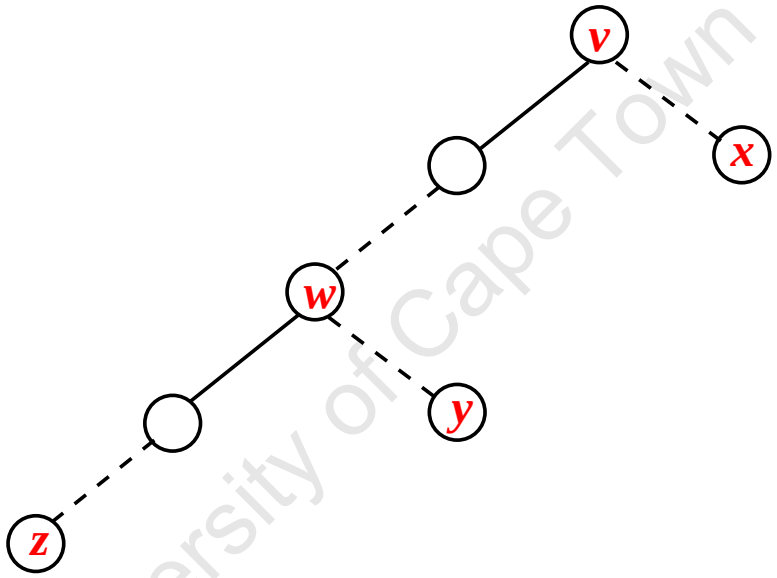


Figure 4.2: General Structure of Minimum Spanning Tree for 3 GO terms x, y and z in the GO DAG. This provides a general representation of 3 GO terms in the GO DAG with a minimum number of edges. - - - means that the branches can go down as low as they can and —— shows the possible existing branches.

The general structure of the minimum spanning tree representing the three GO terms x, y, and z in the GO DAG is given by figure 4.2, other cases are mapped to it. For these three GO terms, the following 6 cases are possible: $IC_T(z) \leq IC_T(y) \leq IC_T(x)$, $IC_T(z) \leq IC_T(x) \leq IC_T(y)$, $IC_T(y) \leq IC_T(z) \leq IC_T(x)$, $IC_T(y) \leq IC_T(x) \leq IC_T(z)$, $IC_T(x) \leq IC_T(z) \leq IC_T(y)$ or $IC_T(x) \leq IC_T(y) \leq IC_T(z)$. Let's consider the first case, if $IC_T(z) \leq IC_T(y) \leq IC_T(x)$ then $\max\{IC_T(x), IC_T(y)\} = IC_T(x)$, $\max\{IC_T(x), IC_T(z)\} = IC_T(x)$, and $\max\{IC_T(y), IC_T(z)\} = IC_T(y)$. From figure 4.2, we have $\frac{IC_T(x,z)}{IC_T(x)} \leq \frac{IC_T(x,y)}{IC_T(x)}$ and as

$\frac{IC_T(y,z)}{IC_T(y)} \leq 1$, it follows that

$$\frac{IC_T(x,z)}{IC_T(x)} + \frac{IC_T(y,z)}{IC_T(y)} \leq 1 + \frac{IC_T(x,y)}{IC_T(x)}$$

Finally putting everything together, we have:

$$\frac{IC_T(x,z)}{\max\{IC_T(x), IC_T(z)\}} + \frac{IC_T(y,z)}{\max\{IC_T(y), IC_T(z)\}} \leq 1 + \frac{IC_T(x,y)}{\max\{IC_T(x), IC_T(y)\}}$$

Meaning that

$$\mathcal{S}_{GO}(x,z) + \mathcal{S}_{GO}(z,y) \leq 1 + \mathcal{S}_{GO}(x,y)$$

The same reasoning can be applied to the other cases.

The quantity $d_{GO}(x,y) = 1 - \mathcal{S}_{GO}(x,y)$ satisfying $0 \leq d_{GO}(x,y) \leq 1$ defines a metric or distance on $\mathcal{N}_{GO}$. The following properties are satisfied:
(i) Positive definiteness: as $0 \leq \mathcal{S}_{GO}(x,y) \leq 1$, we have $1 - \mathcal{S}_{GO}(x,y) \geq 0$ meaning that $d_{GO}(x,y) \geq 0$.
(ii) Symmetry axiom: as $\mathcal{S}_{GO}(x,y) = \mathcal{S}_{GO}(y,x)$, we have $1 - \mathcal{S}_{GO}(x,y) = 1 - \mathcal{S}_{GO}(y,x)$, which means that $d_{GO}(x,y) = d_{GO}(y,x)$.
(iii) Identity axiom: $d_{GO}(x,y) = 0$ implies $1 - \mathcal{S}_{GO}(x,y) = 0$, meaning that $S_{GO}(x,y) = 1$, which implies that $x \overset{\text{GO}}{=} y$.
(iv) Finally, sub-additivity or triangle inequality: from the relation (4.11), we know that $0 \leq \mathcal{S}_{GO}(x,z) + \mathcal{S}_{GO}(z,y) \leq 1 + \mathcal{S}_{GO}(x,y)$. It follows that

$$-[1 + \mathcal{S}_{GO}(x,y)] \leq -[\mathcal{S}_{GO}(x,z) + \mathcal{S}_{GO}(z,y)]$$

Adding 2 on both sides, we have $1 - \mathcal{S}_{GO}(x,y) \leq [1 - \mathcal{S}_{GO}(x,z)] + [1 - \mathcal{S}_{GO}(z,y)]$, which means that $d_{GO}(x,y) \leq d_{GO}(x,z) + d_{GO}(z,y)$.

This metric is referred to as the GO-universal metric, and the more topological information two concepts share, the lower their distance and the more similar or closer they are. Moreover, the similarity formula in (4.10) emphasizes the importance of the shared GO terms by giving more weight to the shared ancestors corrected by the maximum topological information, and thus measuring how similar or close each GO term is to another. Thus, for two GO terms sharing less informative ancestors the distance is higher and the similarity smaller, while for two GO terms sharing more informative ancestors, they are closer and the similarity is higher. Furthermore, the formula in (4.8) shows that the contribution of

a given parent to the term depends on the parent reachability measure. The smaller the reachability measure of that parent and the lower the number of children it possesses, the higher its similarity compared to another parent of the term.

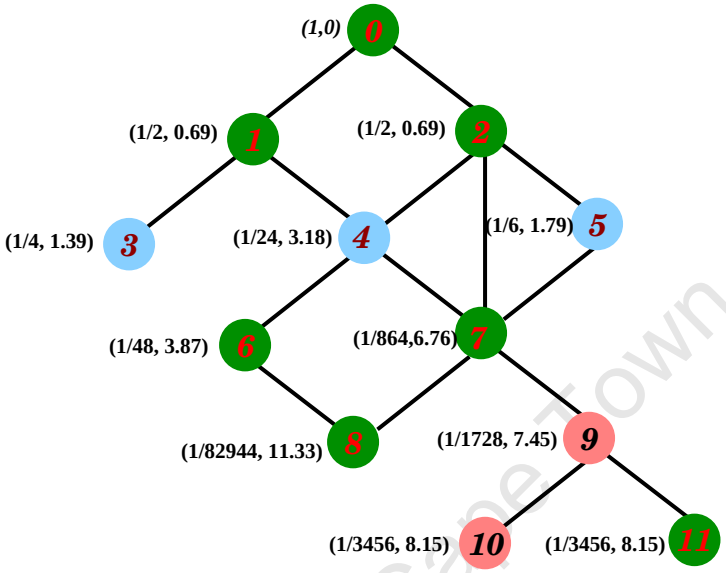


Figure 4.3: *Hierarchical structure illustrating how to compute topological position characteristic and information. Nodes are numbered from* 0 *to* 11 *with* 0 *as a root. The numbers beside each node represent its topological position characteristic and information content.*

To illustrate the way this approach works, consider the hierarchical structure as shown in figure 4.3. In this DAG from top to bottom, we have:

- The topological position characteristic of the root 0, $\mu(0) = 1$, and so its topological information is $\tau(1) = -\ln(1) = 0$.

- As 1 and 2 have only parent 0, which has only these two children with $\mu(0) = 1$, this yields: $\mu(1) = 1/2 = \mu(2)$, and so their topological information is $\tau(1) = -\ln(1/2) = 0.69315 = \tau(2)$.

- 3 has only one direct parent 1, with $\mu(1) = 1/2$ and two children, we have: $\mu(3) = 1/4$ and its topological information is then $\tau(3) = -\ln(1/4) = 1.38639$.

- 4 has two direct parents 1 and 2. 1 has two children with $\mu(1) = 1/2$ and 2 has three children with $\mu(2) = 1/2$. Thus, its topological position characteristic is the product of topological characteristic position of its parents respectively divided by the number of children for each parent $\mu(4) = (1/4) * (1/6) = 1/24$ and its topological information is $\tau(4) = -\ln(1/24) = 3.17806$.

- 5 has only one direct parent 2, which has three children and $\mu(2) = 1/2$. Its topological position characteristic is $\mu(5) = 1/6$ and its topological information is $\tau = -\ln(1/6) = 1.79176$.
Unlike edge-based approaches where nodes and edges are uniformly distributed, and edges at the same level of the ontology correspond to the same semantic distance between terms [191], in this new approach these parameters depend on the topological position characteristic of terms, which are not necessarily the same. In this illustration, nodes 3, 4 and 5 are at the same level but they do not have the same topological position characteristic, thus leading to different topological information or semantic values.

Furthermore, the above illustration reveals that the product in formula (4.8) of topological position characteristic must be carefully taken into account when implementing the approach since the exponential tail-off with increasing depth is severe depending on the density of the hierarchical structure under consideration. Here, we suggest computing $\mu(z)$ iteratively when performing this product and every time the multiplication is done, the obtained value must immediately be converted to a pair of numbers (α, β) such that $\mu(z) = \alpha 10^{\beta}$ with $0.1 \leq \alpha < 1$ and $\beta < 0$. This means that every time the product is performed, the new value is converted to this format so that at the end, the topological position characteristic is just given by (α, β) such that $\mu(z) = \alpha 10^{\beta}$ and $\tau = -\ln(\alpha) - \beta \ln(10)$.

Functional Closeness of Proteins

A given protein may perform several functions, thus requiring several GO terms to describe these functions. For characterized or annotated pair-wise proteins with known GO terms, functional closeness or GO similarities based on their annotations, and consequently the distances between these proteins can be evaluated using the Czekanowski-Dice approach [211] as follows.

$$\mathcal{S}_{\mathcal{F}}(p, q) = \frac{2 \times \left| T^X_{GO}(p) \cap T^X_{GO}(q) \right|}{|T^X_{GO}(p) \cup T^X_{GO}(q)| + |T^X_{GO}(p) \cap T^X_{GO}(q)|} \tag{4.12}$$

where $T^X_{GO}(u)$ is the set of GO terms of a given protein u for a given ontology $X = MF, BP, CC$, and $\left| T^X_{GO}(u) \right|$ stands for its number of elements.

The Czekanowski-Dice's measure is not convenient for use in the case of GO term sets, since GO terms may be similar at some level without being identical. This aspect cannot be captured in the Czekanowski-Dice's measure which only requires the contribution from the GO terms exactly matched between the sets of GO terms of these proteins. One can attempt to avoid this difficulty by incorporating the true path rule in the computation of the intersection and union of GO term sets for proteins. However, in most cases where these proteins are annotated by successive GO terms in the GO DAG, this may lead to the situation where the number of elements in the union of these sets is equal to that of their intersection plus one, in which case, the functional closeness of these proteins is forced to converge to 1, independently of biological contents of the GO terms in the GO DAG. Indeed, let's consider two proteins p and q annotated respectively, with terms 9 and 10, as shown in the fictitious hierarchical structure depicted by figure 4.3.
We have $T^F_{GO}(p) = \{9\}$ and $T^F_{GO}(q) = \{10\}$, and considering an exact match, this will, in general, lead to $\mathcal{S}_{\mathcal{F}}(p,q) = 0$ since $T^F_{GO}(p) \cap T^F_{GO}(q)$ is empty. Furthermore, using the 'true path rule', we have $T^F_{GO}(p) = \{9,7,5,4,2,1,0\}$ and $T^F_{GO}(q) = \{10,9,7,5,4,2,1,0\}$. Thus $T^F_{GO}(p) \cap T^F_{GO}(q) = T^F_{GO}(p)$ with 7 elements and $T^F_{GO}(p) \cup T^F_{GO}(q) = T^F_{GO}(q)$ with 8 elements. Generally, for this case $|T^F_{GO}(p)| = n$, $|T^F_{GO}(q)| = n+1$ will yield $|T^F_{GO}(p) \cap T^F_{GO}(q)| = n$ and $|T^F_{GO}(p) \cup T^F_{GO}(q)| = n+1$, *i.e*, the number of elements in the union of these sets is equal to that of their intersection plus one. This leads to the following functional similarity measure:

$$\mathcal{S}_{\mathcal{F}}(p,q) = \frac{2n}{2n+1} = 1 - \frac{1}{2n+1}$$

Meaning that for n becoming large $\mathcal{S}_{\mathcal{F}}(p,q)$ will be forced to converge to 1. This shows Czekanowski-Dice's measure yields the functional closeness measure which is independent of the semantic value of these terms in the GO DAG.

To overcome this problem, we set up a functional similarity between proteins which emphasizes semantic similarity between terms in their sets of GO terms considered to be uniformly distributed. This functional similarity is given by

$$\mathcal{S}_{\mathcal{F}}(p,q) = \frac{1}{2}\left[\frac{1}{\left|T^X_{GO}(p)\right|}\sum_{t\in T^X_{GO}(p)} \mathcal{S}_{GO}\left(t, T^X_{GO}(q)\right) + \frac{1}{\left|T^X_{GO}(q)\right|}\sum_{t\in T^X_{GO}(q)} \mathcal{S}_{GO}\left(t, T^X_{GO}(p)\right)\right] \quad (4.13)$$

where $\mathcal{S}_{GO}\left(t, T_{GO}^{X}(u)\right) = 1 - d_{GO}\left(t, T_{GO}^{X}(u)\right)$, with $d_{GO}\left(t, T_{GO}^{X}(u)\right)$ the distance between a given term t and a set of terms $T_{GO}^{X}(u)$ for a given protein u, mathematically defined as follows:

$$d_{GO}\left(t, T_{GO}^{X}(u)\right) = \min\{d_{GO}(t, s) : s \in T_{GO}^{X}(u)\}$$

Thus, owing to the fact that $d_{GO}(s, t) = 1 - \mathcal{S}_{GO}(t, s)$, we obtain:

$$\mathcal{S}_{GO}\left(t, T_{GO}^{X}(u)\right) = \max\{\mathcal{S}_{GO}(t, s) : s \in T_{GO}^{X}(u)\} \tag{4.14}$$

This shows that the functional closeness formula emphasizes the importance of the shared GO terms by assigning more weight to similarities than differences. Thus, for two proteins that do not share any similar GO terms the functional closeness value is 0, while for two proteins sharing exactly the same set of GO terms, the functional closeness value is 1.

Evaluating the GO-Universal Metric

Given two GO terms, a set of GO terms, or two GO annotated proteins from a specific proteome, equations (4.10) and (4.13) allow computations of semantic similarities between these GO terms and functional closeness measures between sets of GO terms or between annotated proteins, respectively. To illustrate the approach, we use equations (4.7) and (4.8) to compute the reachability measure $\mu(z)$ and topological information measure $\tau(z)$ of GO terms z in the snapshot of the GO molecular function ontology for the term GO:0004003 (*ATP-dependent DNA helicase activity*) shown in figure 4.4.

Results are listed in table 4.2 and as we can observe, the higher the level of the term, *i.e.*, the further it is from the root node, the higher its topological information, meaning that children are more informative or more specific than their parent, and for two GO terms in the same path, the one of the higher level will either be more informative or topologically identical to that of the lower level.

Using equation (4.10), we calculate the semantic similarity between GO terms in figure 4.4 and results are given in table 4.3. From these results, we see that *ATP-dependent DNA*

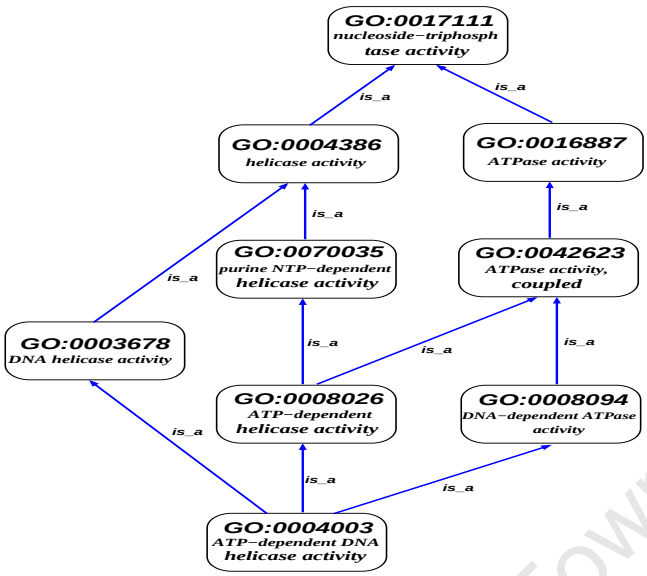


Figure 4.4: *Snapshot of the term GO:0004003 in the molecular function ontology adapted from the sub-GO DAG in the AmiGO browser. This is used to illustrate the effectiveness of the GO-universal metric.*

helicase activity (GO:0004003) is closer, or more similar to *ATP-dependent helicase activity* (GO:0008026) than to *DNA-dependent ATPase activity* (GO:0008094) or to *DNA helicase activity* (GO:0003678). This is numerically due to the influence of *ATP-dependent helicase activity*, reflected by its reachability measure, which is lower than the other terms. It is topologically caused by the higher number of parents the term *ATP-dependent helicase* possesses, some of which are ancestors of *DNA ATPase activity* and *DNA helicase activity*, thus giving the term *ATP-dependent helicase* a higher biological content property than *DNA-dependent ATPase activity* and *DNA helicase activity*.

This illustration shows the benefit of using our approach, as it provides a scalable and consistent measurement method, in which, the semantic similarity of two terms is completely determined by their reachability measures and that of their highest informative ancestor, *i.e.*, the ancestor with the smallest reachability measure. Using the intrinsic topology property of the GO DAG, the semantic similarity measure of two terms is in agreement with the GO consortium vocabulary, in the sense that two terms whose most informative common ancestor is close to the root have smaller shared topological information compared to those having the most informative common ancestor far from the root.

Table 4.2: *Topological position characteristics μ and Information Content τ of GO terms extracted from figure 4.4.*

GO Id	GO Name	Level	$\mu(z)$	$\tau(z)$
GO:0017111	nucleoside-triphosphatase activity	6	0.1861457e-06	1.549674e+01
GO:0004386	helicase activity	7	0.2326821e-07	1.757618e+01
GO:0016887	ATPase activity	7	0.2326821e-07	1.757618e+01
GO:0003678	DNA helicase activity	8	0.5817053e-08	1.896247e+01
GO:0070035	purine NTP-dependent helicase activity	8	0.5817053e-08	1.896247e+01
GO:0042623	ATPase activity, coupled	8	0.1163411e-07	1.826932e+01
GO:0008026	ATP-dependent ATPase activity	9	0.0307619e-16	4.032284e+01
GO:0008094	DNA-dependent ATPase activity	9	0.1057646e-08	2.066722e+01
GO:0004003	ATP-dependent DNA helicase activity	10	0.0026286e-34	8.422920e+01

Table 4.3: *Semantic similarities between pair-wise terms in figure 4.4.*

	GO:0017111	GO:0004386	GO:0016887	GO:0003678	GO:0070035	GO:0042623	GO:0008026	GO:0008094	GO:0004003
GO:0017111	1.00000	0.88169	0.88169	0.81723	0.81723	0.84824	0.38432	0.74982	0.18398
GO:0004386		1.00000	0.88169	0.92689	0.92689	0.84824	0.43586	0.74982	0.20867
GO:0016887			1.00000	0.81723	0.81723	0.96206	0.43586	0.85044	0.20867
GO:0003678				1.00000	0.92689	0.81723	0.43586	0.74982	0.22513
GO:0070035					1.00000	0.81723	0.47027	0.74982	0.22513
GO:0042623						1.00000	0.45308	0.88398	0.21690
GO:0008026							1.00000	0.45308	0.47873
GO:0008094								1.00000	0.24537
GO:0004003									1.00000

As pointed out previously, the biggest limitation of existing approaches based on Resnik's algorithm is that they are constrained by the annotation statistics related to the terms. On the other hand, although Wang's measure is also based on the intrinsic topology of the GO DAG, one of the drawbacks of their approach is that it raises the scalability issue since it requires complete knowledge of the sub-GO DAG of the two terms for which the semantic similarity is being computed and that of all their common ancestors. However, it should be noted that GO is expanding and increasing in size, and term relationships are becoming more and more important. Thus, a semantic similarity measurement approach should be effective independently of the size of the GO DAG.

Another negative aspect of Wang's approach is that it essentially relies on the semantic factors of 'is_a' and 'part_of' relations, which may be user tunable parameters. It is not clear for which values of these semantic factors the semantic similarity measure yields the optimal value of biological content of terms. Moreover, these semantic factors make the similarity value between a given child and its direct parent independent of the number of children that the parent term has, as shown in equation (4.6). Considering the GO

DAG, the semantic similarity between a given term and its child should not only depend on the number of parents the child term possesses, but also on the number of children that the parent term possesses. Thus, the semantic similarity must be sensitive to the number of children of the parent term. The greater the number of children, the smaller the semantic similarity to each of its children in order to map the idea of 'the more the term is used for annotation the lower its semantic value' in semantic value measures based on the annotation statistics. This is actually in agreement with human judgement.

4.2.3 Functional Closeness versus Functional Link Score

In order to validate a given semantic similarity measure, an external set of data, such as sequence similarity, gene expression or protein-protein interactions, correlated with GO annotations are used. Thus, to validate the GO-universal metric, we used the relation between functional link confidence scores in the MTB functional network produced in chapter 3, and their GO annotations downloaded from the Integr8 project of the European Bioinformatics Institute (EBI).

Using the relationship between link confidence scores in the functional network, and similarity scores between their GO annotations as a means of assessing the effectiveness of our measure, we analyzed the distribution of interactions with respect to these two scores for all three ontologies, 'molecular function', 'biological process' and 'cellular component'. In the computation, link confidence and functional similarity scores, ranging from 0 to 1, corresponded well. An increase of GO similarity score corresponds to an increase in the functional link confidence score. Those values close to one indicate high confidence and those close to zero indicate low confidence.

We analyzed the distribution of these functional interactions with respect to their functional link confidence and GO similarity scores for each confidence range and for all the three ontologies using all interaction datasets and results are shown in table 4.4. Note that values in table 4.4 are ratios of occurrences of functional link and functional similarity scores at a given confidence level. At this confidence level, the rate r of minimum and maximum between ratio values of functional link and functional similarity between proteins defines positive or negative relationship between functional link and functional similarity scores,

Table 4.4: *Interaction distribution per confidence type and per ontology for evaluating our approach in terms of percentage of co-occurrence of functional link (Link Score) and our similarity scores (GO Sim) at three levels of confidence: low (less than* 0.3*), medium (between 0.*3 *and* 0.7*) and high (greater than* 0.7*).*

	Low confidence		Medium confidence		High confidence	
Name Space	Link Score	GO Sim	Link Score	GO Sim	Link Score	GO Sim
CC	0.09956	0.20391	0.45643	0.18710	0.44401	0.60899
MF	0.04788	0.13826	0.42159	0.30445	0.53053	0.55729
BP	0.05305	0.20719	0.39745	0.25652	0.54950	0.53629

according to the fact that $r \geq 0.5$ or $r < 0.5$.

These results show that there is a positive relationship between link confidence and similarity scores from the medium to high confidence levels for the MF and BP ontologies, indicating that where link scores and GO similarity measures are greater than 0.3 we are confident that where proteins are functionally linked, these proteins are also more likely to be annotated to similar GO terms. At a high confidence level, functional link and GO similarity are related at the level of $(0.53053/0.55729) \times 100 = 95\%$ and $(0.53629/0.54950) \times 100 = 98\%$ for the MF and BP ontologies, respectively. Similarly, at a medium level they are related at the level of 72% and 65%, respectively. If we consider both medium and high levels, *i.e.*, we look at the occurrences of functional link and GO similarity between proteins at a confidence level greater than 0.3, then functional link and GO similarity are correlated at 91% and 84% for the MF and BP ontologies, respectively, indicating that linked proteins tend to have similar molecular functions and also to be involved in similar biological processes when we have medium to high confidence in their functional link.

As previously pointed out (in chapter 3), some data (e.g. sequence data) are better for predicting protein molecular functions, while other data are better for biological process. Thus, we have conducted the same analysis using only sequence data and results are shown in Table 4.5. Once again these results indicate that there is a positive relation between link confidence score and GO closeness at a medium and high confidence level, indicating that where link score and GO closeness measures are greater than 0.3 we are confident (confidence or reliability > 0.7) that where proteins are functionally linked, these proteins are also more likely to be annotated to similar GO terms, especially for 'molecular function'

Table 4.5: *Interaction distribution per confidence type and per ontology for evaluating our approach in terms of percentage of co-occurrence of functional links from sequence data and our GO similarity scores at three levels of confidence: low (less than* 0.3*), medium (between 0.*3 *and* 0.7*) and high (greater than* 0.7*).*

	Low confidence		Medium confidence		High confidence	
Name Space	Link Score	GO Sim	Link Score	GO Sim	Link Score	GO Sim
CC	0.00596	0.16520	0.57302	0.16390	0.42103	0.67089
MF	0.00220	0.02845	0.44469	0.25536	0.55311	0.71618
BP	0.00141	0.08794	0.42060	0.27597	0.57799	0.63609

at $(0.55311/0.71618)\times100 = 77\%$ and 'biological process' at $(0.57799/0.63609)\times100 = 91\%$ at high confidence, and 58% and 66% at medium confidence for molecular function and biological process, respectively. This indicates that homologous proteins tends to have similar molecular functions and are also involved in similar biological processes from a medium level of confidence and upwards.

4.3 Annotation Prediction Algorithm

There are several goals that a function prediction algorithm needs to meet in the current genomic era. These include improvement of annotation quality and genomic coverage, *i.e.*, increase the proportion of genes or gene products in a genome which are annotated [212]. Despite the high degree of noise that interaction data from high throughput experiments contain, making them potentially unreliable, uncontested successes have been recorded from the use of computational approaches to predict functions of uncharacterized data by using these data. Several approaches have been proposed for predicting protein functions from functional networks and are mainly classified into two categories, namely global network topology and local neighborhood based approaches. Global network topology based approaches use global optimization [213, 214, 215], probabilistic methods [216, 217, 218, 219] or machine learning [220, 221, 222] to improve the prediction accuracy using the global structure of the network under consideration. In the case of local neighborhood based approaches, known as 'Guilt-by-Association', 'Majority Voting' or 'Neighbor Counting' [223], direct interacting neighbors of proteins are used to predict protein functions.

The dualism of "Guilt-by-Association" and "Global" prediction approaches for characterizing a protein has raised a debate separating the Bioinformatics community into divergent groups with different views concerning how the annotation prediction of proteins labelled "uncharacterized" should be engineered. On one hand, there are proponents of the "Guilt-by-Association" strategy, stating that a gene or gene product shares the function of the most closely related genes of known functions, thus predicting protein functions by observing the patterns of each protein's neighborhood. This fraction highlights the inability of global prediction approaches to provide significant improvement over the simple and elegant local prediction approach [224]. On the other hand, the advocates of the "Global" prediction approach argue for a global view of the protein-protein interaction network to achieve efficient annotation prediction.

From the computational side, the "Guilt-by-Association" prediction approach may be an excellent and more straightforward approach since the "Global" prediction approach raises a scalability issue for large datasets which may not be proportional to the prediction improvement. However, this straightforward approach may lead to systematic error especially in the case where the protein under consideration does not share functions with any of its direct neighbors. This case was depicted in the Yeast Proteome Network and therefore, Jin and Cho [225] proposed a new approach for dealing with this type of local protein association behavior by building a "Protein Interaction Network Dictionary (PIND)" in which the protein target's function is obtained from characterized proteins whose direct interacting neighbors shared a certain level of similarity with the protein's target interacting neighbors. This phenomenon has also been demonstrated by Chua et al. [226, 227], who showed that in many cases proteins share functional similarity with level-2 neighbors, and level-2 neighbors have an above average likelihood of sharing functional similarity. They introduced a functional similarity weight (FS-Weight) method for predicting protein functions from protein interaction data using level-1 and level-2 neighbors. Here, we present an annotation prediction model which uses direct interacting neighbors combined with second level interacting neighbors to achieve efficient trade-off between the scalability issue, prediction improvement and genomic coverage.

4.3.1 Guilt-by-Association Approach

The "Guilt-by-Association" approach consists of assigning to a protein the function that occurs most frequently among its direct interacting partners [223]. Let $\mathcal{N}$ be the set of all proteins in the proteome of the organism under study, $T^F_{GO}(p)$ the set of functional GO terms of a given protein $p \in \mathcal{N}$, and $\mathcal{F}_{GO} = \bigcup_{p \in \mathcal{N}} T^F_{GO}(p)$ the set of all functional GO terms for the proteome of the organism. The frequency $f_t(p)$ of the term $t \in \mathcal{F}_{GO}$ occurring among direct interacting partners of protein $p \in \mathcal{N}$ is computed as

$$f_t(p) = \sum_{q \in \mathcal{N}_p} \delta_q(t) \tag{4.15}$$

where $\mathcal{N}_p$ refers to the set of all direct interacting partners of protein p, and δ_q is the $q-$function indicator given by

$$\delta_q(t) = \begin{cases} 1 & \text{if the protein } q \text{ performs the function } t \\ 0 & \text{otherwise.} \end{cases}$$

In this case, the frequency $f_t(p)$ is actually the number of interacting partners of protein p performing the function t, and serves as a score $\mathcal{S}_t(p)$ of function t for protein p, and the n most frequent functions, *i.e.*, the n functions with the highest score among interacting partners of protein p are assigned as the n most likely functions for the protein p.

The score of a function or functional GO term x for a protein p as set by the Majority Voting approach does not take into account the number of proteins annotated by the term x across the entire functional network. Indeed, the more proteins that are annotated by the term x, the more likely the term is to be found among the interacting partners of protein p, and also the more likely the protein p may be annotated by the term x. Thus, if two terms occur at equal frequency among p interacting partners the one annotating more proteins will obviously have a higher probability of being found in p. However, the Majority Voting would assign the same score to both functions. A variant of the Majority Voting, the Chi-Square approach, has been proposed [228] in order to overcome the above limitation. The Chi-Square approach assigns n largest Chi-Square score functions to the protein p, calculated as [229]

$$\mathcal{S}_t(p) = \frac{[f_t(p) - \epsilon_t(p)]^2}{\epsilon_t(p)} \tag{4.16}$$

where $f_t(p)$ is defined in equation (4.15), and $\epsilon_t(p) = n \times \pi_t$, the expected number of partners performing the function t, with π_t the fraction of proteins performing function t among all the proteins and n the total number of proteins in the network.

Another shortcoming of the "Guilt-by-Association" approach beyond the fact that it is bound to fail in the case where the target protein's neighbors are also uncharacterized, is that it does not consider the structure of the entire ontology used to predict these functions, *i.e.*, it does not take into account the relationships between all annotations in the ontology under consideration. In most cases, the level at which a term can be considered to be specific or informative in the annotation hierarchy is fixed, thus leading to partial coverage of the annotation structure which compromises the prediction analysis since terms used are not comparable. Furthermore, the "Guilt-by-Association" approach provides high confident predictions only to proteins sharing significantly similar functions with their interacting neighbors. This means that these approaches may be misleading and can yield a very high false positive rate [230]. Taking into account all these observations, we are predicting the MTB uncharacterized proteins, where possible, by using the functional organization of level-1 and level-2 interacting partners of the protein under consideration in the functional network using the MF and BP ontologies of the GO DAG. Throughout this annotation prediction process, instead of using exact matches only, relationships between GO terms in the GO DAG structure are considered through the GO term's semantic similarity defined in section 4.1. This combination is expected to improve the prediction quality and the genome coverage.

4.3.2 Protein Function Prediction

We are exploiting the underlying biological principle of the functional network structure and its dynamics which allow the system to be stable and robust, functioning in a reliable way. More precisely, we observe the level 1 and 2 neighbors' function occurrence patterns to identify the key principles driving the functions imposed to a protein by its neighbors referred to as "traces" of the underlying biological organization of the system. Depending

on the features of the protein under consideration obtained from its direct and level-2 interacting partners, the optimal strategy, which consists of finding the best use of 'traces' of underlying biological principles, is applied to predict more accurately the functions of the protein. This is described in figure 4.5.

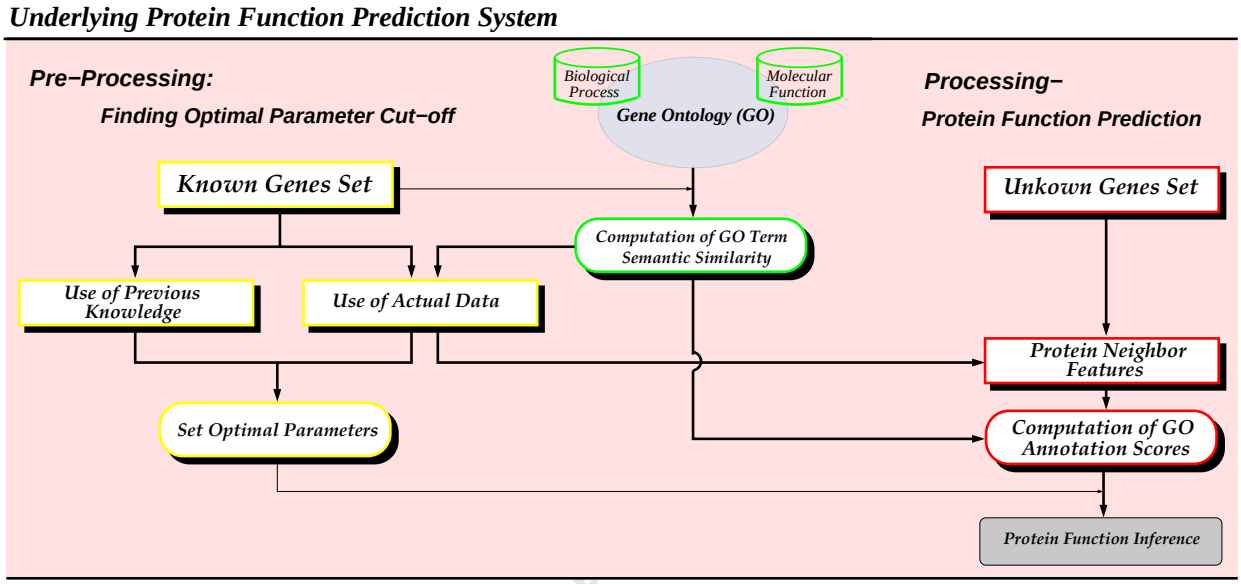


Figure 4.5: *Protein function prediction system flow diagram.*

As we are using GO terms, namely GO molecular function and biological process terms, to predict functions of uncharacterized proteins, we must take into account the issue related to the GO DAG structure in which two terms can be similar without being identical. Thus, the annotation score of a GO term requiring an exact match as defined previously in formulas 4.15 and 4.16 is not applicable. We set up a new annotation score using the GO universal similarity metric for predicting the most likely functions carried out by proteins. This annotation score can be applied to known proteins or genes in the functional network to determine the optimal cut-off from which a term can be considered to be a function of the protein under consideration. This cut-off can also be determined in advance using previous annotation knowledge about the organism under study.

GO Annotation Score and Prediction Algorithm

As pointed out previously, the GO annotation score must take into account the frequency of GO terms occurring among protein interacting partners and semantic similarity between them. Thus, the GO annotation score of the term t for a protein p is given by

$$\mathcal{S}_t^{GO}(p) = \frac{1}{|T_{GO}(p)|} \sum_{s \in T_{GO}(p)} \mathcal{S}_{GO}(t, s) \tag{4.17}$$

where $T_{GO}(p) = \overset{+}{\underset{q \in \mathcal{N}_p}{\cup}} T_{GO}^X(q)$ is the multiset or bag (set with possible multiple copies of an element allowed) of all the GO terms occurring among interacting partners of protein p, $\mathcal{S}_{GO}(t, s)$ is the semantic similarity between two GO terms t and s defined in (4.10), and $T_{GO}^X(q)$ is the set of GO terms of a given protein q for an ontology $X = MF$ or BP. As in the standard Guilt-by-Association approach, this score estimates the likelihood of the protein p being annotated by the GO term t.

In fact, the formula in (4.17) depends on number of times a given GO term occurs among protein neighbors. This scoring formula emphasizes the importance of the GO terms' frequency of occurrence among the interacting partners of the protein under consideration and it is clear that this score increases with the frequency of occurrence of a term, since one can easily show that for a given positive real number ϵ and integers n and k such that $n \geq k \geq 1$, we have

$$\frac{\epsilon}{n} \leq \frac{\epsilon k}{n - 1 + k}$$

This indicates that the scoring formula considers the semantic similarity of GO terms in the GO DAG and takes into account the GO terms' frequency of occurrence by increasing the weight of the terms occurring several times among protein neighbors.

The GO annotation score is computed for every term occurring among protein neighbors. The threshold is set based on previous knowledge about data or calculated using data on known protein functions, and GO term exceeding this threshold is assigned as a function of the protein under consideration.

Evaluation of the Prediction Approach

In order to derive the traces of the underlying biological organization of the MTB functional network, we make use of this network to determine the optimal parameter cut-off that provides the best quality of predictions when using the following approaches:

- Guilt-by-Association approach using direct interacting neighbors of the protein target, with the score computed as in the equation (4.17). This approach is referred to as the GO-GA approach from now on.

- Guilt-by-Association approach derived from direct interacting neighbors of other proteins. This is basically the PIND approach, where the potential functions of the protein target are functions of other proteins whose direct interacting neighbors share significant similarity with the interacting neighbors of the protein target. In this case, the functional similarity between the $\mathcal{N}_p$ and $\mathcal{N}_q$ set of neighbors to proteins p and q, respectively, is given by $\mathcal{S}_{\mathcal{F}}(T_p, T_q)$ defined in equation (4.13) with $T_u = \overset{+}{\bigcup}_{v \in \mathcal{N}_u} T^X_{GO}(v)$ as the multiset or bag of all the GO terms occurring among interacting partners of protein u. The potential functions of the protein target is the multiset comprised of all the functions occurring among proteins whose interacting neighbors share functional similarity with the protein target's neighbors. This is referred to as the GO-PIND approach.

- The method that combines the GO-GA and GO-PIND approaches, referred to as the GO-GAPIND-1 approach. In this method, the potential functions of the protein target are obtained by merging the functions occurring among its direct interacting partners and those of all the proteins whose interacting neighbors share functional similarity with the protein target's neighbors.

- Guilt-by-Association approach which uses the level-1 and level-2 neighbors of the protein target. Investigating the relation between interacting neighbors of a given protein using network topology, Hua et al. [226, 227] show that in many cases, a protein shares functional similarity with level-2 neighbors (2 branch-lengths away), which may be used to provide greater coverage during function inference. However, these level-2 neighbors yield high false positives and they introduced a Functional Similarity Weight (FS-Weight), a topological measure from which both level-1 and level-2 neighbors that are more likely to share functions with the protein target are identified in order to enhance the quality of

predictions. They have considered the transitivity of this topological functional similarity indicating that if a protein p is topologically similar to a protein u, and protein u is topologically similar to protein q, then proteins p and q may show some degree of topological similarity. Thus, the transitive FS-weight measure is given by

$$\mathcal{S}_{TR}(p,q) = \max\left(\mathcal{S}_{FS}(p,q), \max\left\{\mathcal{S}_{FS}(p,u) * \mathcal{S}_{FS}(u,q) : u \in \mathcal{N}_q^+\right\}\right) \tag{4.18}$$

where $\mathcal{S}_{FS}(p,q)$ is the FS-weight score between proteins p and q in the functional network, defined as

$$\mathcal{S}_{FS}(p,q) = \frac{2\left|\mathcal{N}_p^+ \cap \mathcal{N}_q^+\right|}{\left|\mathcal{N}_p^+ - \mathcal{N}_q^+\right| + 2\left|\mathcal{N}_p^+ \cap \mathcal{N}_q^+\right| + \lambda_{pq}} \times \frac{2\left|\mathcal{N}_p^+ \cap \mathcal{N}_q^+\right|}{\left|\mathcal{N}_q^+ - \mathcal{N}_p^+\right| + 2\left|\mathcal{N}_p^+ \cap \mathcal{N}_q^+\right| + \lambda_{qp}} \tag{4.19}$$

with $\mathcal{N}_x^+$ the set containing x and its direct neighbors, and λ_{xy} given by

$$\lambda_{xy} = \max\left(0, n_{avg} - \left(\left|\mathcal{N}_x^+ - \mathcal{N}_y^+\right| + \left|\mathcal{N}_y^+ - \mathcal{N}_x^+\right|\right)\right) \tag{4.20}$$

where n_{avg} is the average number of direct neighbors, which is in fact the average degree of the functional network, given by $n_{avg} = 2 \times |\mathcal{L}| / |\mathcal{N}|$, with $|\mathcal{N}|$ and $|\mathcal{L}|$ being the number of proteins and functional links, respectively, in the functional network.

Considering the impact of topological similarity weight, the GO annotation score of a term t for a protein p is computed as follows:

$$\mathcal{S}_t^{GO}(p) = \frac{1}{Z_p} \sum_{q \in \mathcal{N}_p} \left(\mathcal{S}_{TR}(p,q)\, \mathcal{S}_{GO}\left(t, T_{GO}^X(q)\right) + \sum_{u \in \mathcal{N}_q^*} \mathcal{S}_{TR}(q,u)\, \mathcal{S}_{GO}\left(t, T_{GO}^X(u)\right)\right) \tag{4.21}$$

where $\mathcal{N}_q^*$ is the set of interacting partners of q different from the protein target p, and $\mathcal{S}_{GO}\left(t, T_{GO}^X(v)\right)$ the GO semantic similarity score between a term t and a protein v defined in equation (4.14), which is set to zero if v is uncharacterized, and Z_p is a normalization factor given by

$$Z_p = \sum_{q \in \mathcal{N}_p} \left(\mathcal{S}_{TR}(p, q) + \sum_{u \in \mathcal{N}_q^*} \mathcal{S}_{TR}(q, u) \right) \tag{4.22}$$

The formula (4.21) above is adapted from [227] to consider the structure of the GO DAG. The conventional method only considers the exact match between terms, ignoring the overall GO DAG structure and relationship between terms. This GO annotation score is set to zero if $Z_p = 0$, and similarly to the GO-GA approach, this annotation score estimates the confidence level of annotating the protein p by the term t. This approach is referred to as the GO-FS approach. Note that the level-2 interacting partners are included in the computation of the GO annotation score and each occurrence of a protein is considered. This means that a characterized level-2 interacting partner is included as many times as the number of different level-1 neighbors it interacts with. Furthermore, level-1 characterized interacting partners that are level-2 neighbors will contribute more to this score.

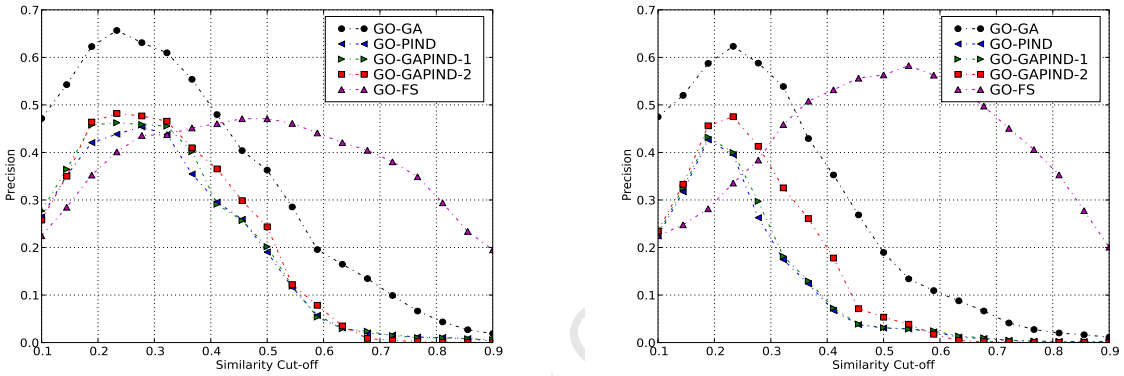
- Combining GO-GAPIND-1 and also considering the level-2 interacting proteins is referred to as the GO-GAPIND-2 approach. In this case, the set of potential functions of the protein target is obtained from level-1 and level-2 neighbors, and from proteins whose level-1 neighbors share significant GO-functional similarity with level-1 neighbors of the protein target. The GO annotation score is computed in the same way as in the GO-GA approach. Note that a given level-2 interacting partner contributes to the GO-annotation score as many times as it occurs as a partner of level-1 proteins, and the level-1 that is also a level-2 partner contributes to the score as a level-1 as well as a level-2 partner.

1. Optimal Parameter Estimation

To determine the optimal GO annotation score threshold from which a GO term may be more likely to be a protein target's annotation, we vary this metric from 0.1 to 0.9 and extract the values that produce the best prediction for each approach described above. We are considering proteins to be functionally similar if their functional similarity score from the GO semantic similarity is greater than 0.4, and the same cut-off is used for semantic similarity between GO terms. This is due to the results displayed in table 4.4 and table 4.5, which show that proteins' functional similarity are positively related to functional link scores from the medium confidence level ($\geq$ 0.3) and upwards in these functional interaction networks.

Table 4.6: *GO annotation score threshold of each of the five approaches and the corresponding highest precision achieved.*

	MF ontology		BP ontology	
Approach	Threshold	Precision	Threshold	Precision
GO-GA	0.23	0.623	0.23	0.657
GO-PIND	0.19	0.427	0.28	0.454
GO-GAPIND-1	0.19	0.432	0.23	0.463
GO-FS	0.54	0.475	0.50	0.482
GO-GAPIND-2	0.19	0.583	0.23	0.471



(a) Precision variation of the five function prediction approaches for BP ontology.

(b) Precision variation of the five function prediction approaches for MF ontology.

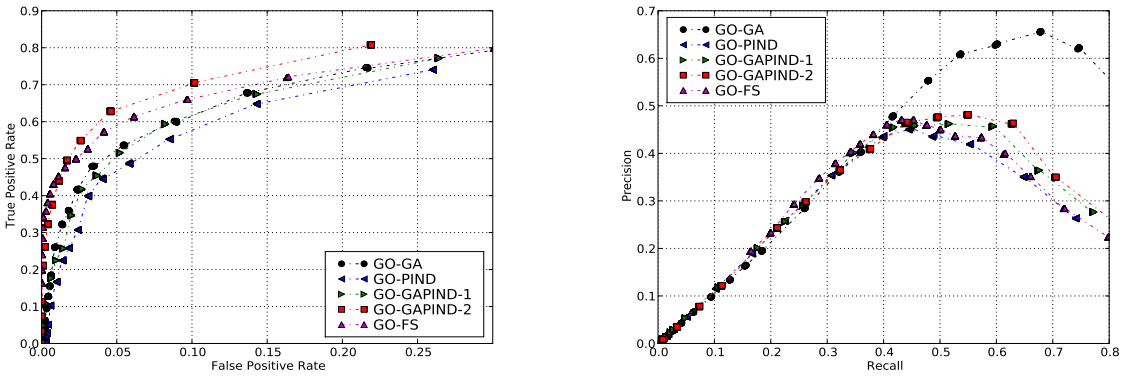
Figure 4.6: *Precision analysis to determine the optimal GO annotation score cut-off.*

For deriving the optimal GO annotation score threshold, we used one of the most important measures of the quality of a prediction approach, the precision of prediction, which measures the proportion of GO terms correctly predicted for a protein among all protein's actual GO terms, given by

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.23}$$

where TP (true positive) represents the number of GO terms correctly predicted, *i.e.*, the number of observed GO terms which are similar (in terms of GO semantic similarity) as those predicted, and FP (false positive) is the number of predicted GO terms different from the actual protein's GO terms.

Results on precision variation in terms of GO annotation score cut-off are given in figure 4.6



(a) ROC curves.

(b) P-ROC curves.

Figure 4.7: *Performance analysis of the function prediction approaches for BP ontology.*

and the optimal cut-off inferred from each approach with the corresponding highest precision are given in table 4.6.

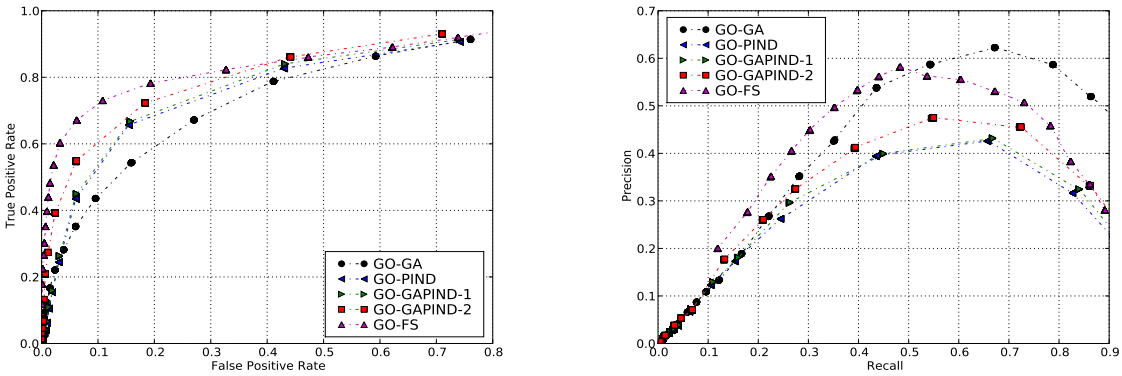
2. Performance of Prediction Approaches

Five different protein function prediction approaches described in the section 4.3.2 have been compared using Receiver Operating Characteristic (ROC) [231, 232] and Precision-Recall Operating Characteristic (P-ROC) [233] curve analyses. We used leave-one-out cross-validation to evaluate the the performance of these prediction approaches computing the false positive (1 − specificity) and true positive rate (sensitivity or recall), given respectively by

$$1 - \text{specificity} = \frac{FP}{FP + TN} \quad \text{and} \quad \text{sensitivity} = \frac{TP}{TP + FN}$$

where TN and FN are true negative and false negative, respectively.

The ROC curve plots false positive rate against true positive rate and P-ROC curve plots precision against recall. In order to compare the performance of these approaches, we combined their related ROC and P-ROC curves and results are shown in figure 4.7 and figure 4.8 for the BP and MF ontologies, respectively. With the exception of the GO-GA approach, which always provides better precision than the others, these results indicate that the GO-GAPIND-2 approach performs better than other approaches for the BP ontology



(a) ROC curves.

(b) P-ROC curves.

Figure 4.8: *Performance analysis of the function prediction approaches for MF ontology.*

and the GO-FS approach is better than the others when dealing with the MF ontology. This means that the GO-GA approach yields the best quality prediction, and as approaches using level-2 interacting neighbors perform better in terms of ROC curves they can be used to complement the GO-GA approach to improve genome annotation coverage. For this purpose, as GO-GAPIND-2 performs better for the BP ontology and GO-FS for the MF ontology, GO-GAPIND-2 is applied when predicting protein annotations with GO BP terms and the GO-FS approach is used for prediction of GO MF terms.

4.4 Annotation Prediction of MTB Proteome

We have used results of the ROC and P-ROC curves for the five different protein function prediction approaches depicted in figures 4.7 and 4.8, and shown that the GO-GA approach yields better quality annotations, and the GO-GAPIND-2 and GO-FS approaches perform better than others for the BP and MF ontologies, respectively.

Algorithm 1 MTB Protein Function Prediction Algorithm.

Input: Functional Network, GO Term Semantic Similarity, Transitive FS-Weight Measures, and GO-annotation Score Cut-off

Output: Predicted Functions of Uncharacterized Proteins

```
for each uncharacterized protein p do
  if p has characterized level-1 neighbors then
    run GO-GA approach
  else
    if p has characterized level-2 neighbors then
      if ontology := MF then
        run GO-FS approach
      else if ontology := BP then
        run GO-GAPIND-2 approach
      end if
    else
      Protein functions cannot be inferred using current data.
    end if
  end if
end for
```


We therefore ran the prediction algorithm, which uses the GO-GA approach complemented by the GO-GAPIND-2 or GO-FS approach, as described in algorithm 1, to predict functions of uncharacterized proteins in the MTB proteome.

To illustrate the effectiveness of the algorithm used, we apply the algorithm to some proteins with known functions, comparing their true functions to the predicted ones. True and predicted functions with their GO functional similarity for the selected set of proteins are shown in table 4.7.

For uncharacterized proteins in the MTB proteome, the functions of 1930 proteins have been predicted using the BP ontology and 1590 proteins using the MF ontology, representing, respectively, 83% and 76% of the total number of uncharacterized proteins in the MTB proteome. This has been achieved using an optimal cut-off of 0.23 for the BP and MF ontologies under the GO-GA approach, and 0.23 and 0.54 for the BF and MF ontologies under the GO-GAPIND-2 and GO-FS approaches, respectively. The prediction approach used was unable to predict functions for 391 proteins for the BP ontology and 497 proteins for the MF ontology due to the fact that these proteins are clustered with uncharacterized proteins. This represents approximately 9% and 12% of the proteome which

Table 4.7: *True functions of some characterized proteins and their predicted functions using algorithm 1. The top of the table is for BP and the bottom for MF.*

UniProt-Acc	Gene-Name	True Functions		Predicted Functions			GO Similarity
		GO-ID	GO name	GO-ID	[GO name]	Annot. score	
P64259	mraY	GO:0051301	cell division	GO:0051301		0.24148	0.71425
		GO:0007049	cell cycle	GO:0007049		0.23762	
		GO:0007047	cell wall organization				
		GO:0008360	regulation of cell shape				
		GO:0009252	peptidoglycan biosynthetic process				
P96203	ppsD	GO:0008152	metabolic process	GO:0008152		0.55361	0.83313
				GO:0055114	oxidation reduction	0.48297	
		GO:0009058	biosynthetic process	GO:0009058		0.40118	
				GO:0006629	lipid metabolic process	0.32551	
				GO:0005975	carbohydrate metabolic process	0.32206	
Q8VJ85	MT3107	GO:0006313	transposition, DNA-mediated	GO:0006313		0.70999	0.72589
				GO:0008033	tRNA processing	0.32890	
				GO:0006400	tRNA modification	0.27794	
O53730	sigK	GO:0006355	regulation of transcription, DNA-dependent	GO:0006355		0.40269	1.00000
		GO:0045449	regulation of transcription	GO:0045449		0.39618	
		GO:0006350	transcription	GO:0006350		0.36396	
		GO:0006352	transcription initiation	GO:0006352		0.36345	
O05577	moeA1	GO:0006777	Mo-molybdopterin cofactor biosynthetic process	GO:0006777		0.54337	0.83454
		GO:0032324	molybdopterin cofactor biosynthetic process	GO:0032324		0.38634	
				GO:0008152	metabolic process	0.23470	
P0A595	glbO	GO:0006810	transport	GO:0006810		0.48071	1.00000
		GO:0015671	oxygen transport	GO:0015671		0.48071	
P0A590	glnA1	GO:0006807	nitrogen compound metabolic process	GO:0006807		0.24062	0.77629
		GO:0009399	nitrogen fixation				
		GO:0006542	glutamine biosynthetic process				
P67298	Rv1960c	GO:0045449	regulation of transcription	GO:0045449		1.00000	1.00000
Q7D9L7	MT0605	GO:0003677	DNA binding	GO:0003677		0.52253	0.96406
		GO:0003700	transcription factor activity	GO:0003700		0.48369	
				GO:0043565	sequence-specific DNA binding	0.45706	
Q7D638	MT3196	GO:0003824	catalytic activity	GO:0003824		0.32775	0.66761
		GO:0016787	hydrolase activity	GO:0016787		0.33174	
		GO:0008270	zinc ion binding				
				GO:0016491	oxidoreductase activity	0.24373	
				GO:0016829	lyase activity	0.23453	
				GO:0004803	transposase activity	0.23027	
O53461	MT1814	GO:0004519	endonuclease activity	GO:0004519		0.54545	1.00000
		GO:0003676	nucleic acid binding	GO:0003676		0.45455	
O06831	fadD12	GO:0003824	catalytic activity	GO:0003824		0.54984	0.82872
		GO:0016874	ligase activity	GO:0016874		0.51981	
				GO:0016740	transferase activity	0.34454	
				GO:0004467	long-chain-fatty-acid-CoA ligase activity	0.29613	
				GO:0016788	hydrolase activity, acting on ester bonds	0.23986	
P72047	rfbE	GO:0005524	ATP binding	GO:0005524		0.34172	1.00000
		GO:0000166	nucleotide binding	GO:0000166		0.27270	
		GO:0017111	nucleoside-triphosphatase activity	GO:0017111		0.23578	
		GO:0016887	ATPase activity	GO:0016887		0.23484	

are uncharacterized with respect to the BP and MF ontologies, respectively.

4.5 Novel Decryption of the MTB Genome Biology

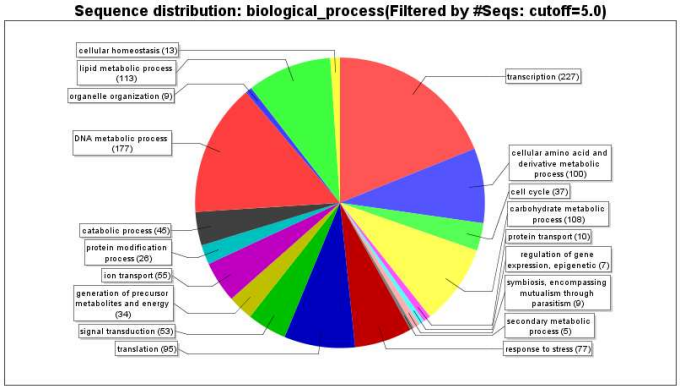
Available data shows 4195 protein coding genes in the MTB CDC1551 genome, with a total of 1874 and 2108 proteins which are characterized with respect to the BP and MF ontologies, respectively. The extensive integration of biological interaction data from different sources has yielded a functional network containing 4136 proteins. We ran the annotation prediction algorithm on the MTB network produced to predict, where possible, functions of proteins labelled 'uncharacterized'. The resulting dataset consists of a total of 3804 proteins with predicted functions using the GO biological process ontology and 3698 proteins have predicted functions with respect to the GO molecular function ontology. The analysis of these data show that out of 3804 proteins with predicted functions with respect to BP terms, 1343 proteins representing approximately 35% are involved in metabolic processes, and metabolic process in itself covers about 15% over all processes occurring in the system, as shown in table 4.8. Note that a protein may be involved in multiple processes, and a general view of the most common MTB CDC1551 proteome processes is shown in table 4.8.

Further analyses of the GO terms predicted for the uncharacterized proteins were done using the BLAST2GO program (`http://blast2go.bioinfo.cipf.es/start_blast2go`), which uses the Fishers Exact Test for over-representation analysis. Table 4.9 and the charts in Figure 4.9 show that the majority of newly predicted functions for uncharacterized proteins include basic cellular processes such as transcription and translation, as well as lipid metabolism. In terms of level of GO term predicted, for MF, most terms are at level 3, while the BP terms predicted go down to level 9, with the majority at level 6, showing that some of the newly predicted terms may be reasonably specific.

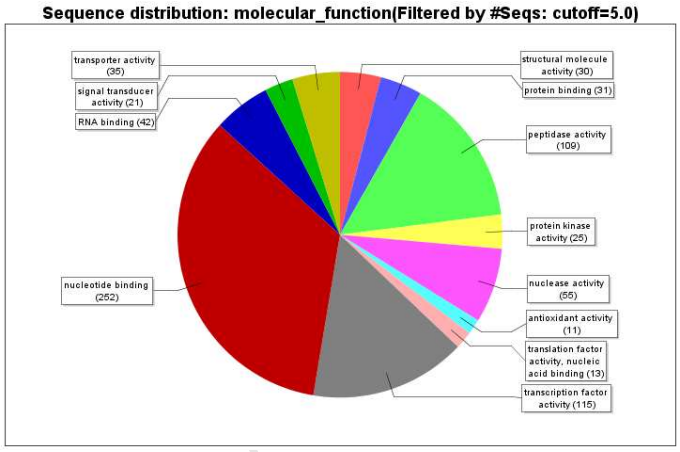
Of particular importance in the proteome are proteins of the PE/PPE family, which are generally uncharacterized, but suspected to be involved in antigenic variability. Functions of a total of 142 out of 147 proteins of this family have been predicted, and some of these predictions are shown in table 4.10.

Table 4.8: *Different processes in which MTB proteins are mostly involved.*

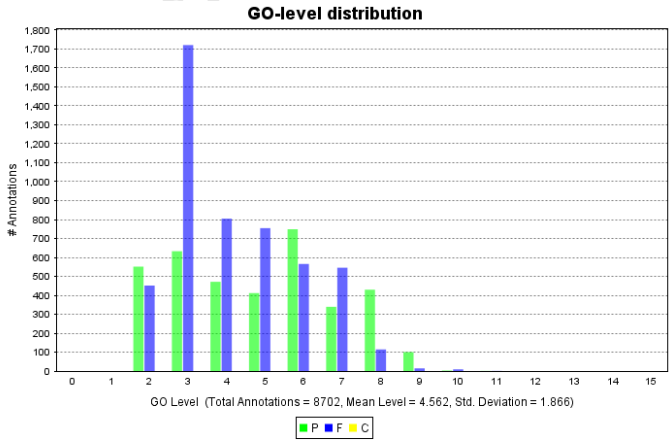
GO-ID Process	GO name	Number of Proteins in	Weight
GO:0008152	metabolic process	1343	14.96
GO:0055114	oxidation reduction	935	10.42
GO:0006355	regulation of transcription, DNA-dependent	379	4.22
GO:0045449	regulation of transcription	352	3.92
GO:0006810	transport	331	3.69
GO:0006350	transcription	316	3.52
GO:0006412	translation	226	2.52
GO:0006508	proteolysis	215	2.40
GO:0009058	biosynthetic process	209	2.33
GO:0055085	transmembrane transport	162	1.81
GO:0006629	lipid metabolic process	150	1.67
GO:0005975	carbohydrate metabolic process	149	1.66
GO:0006281	DNA repair	144	1.60
GO:0008652	cellular amino acid biosynthetic process	128	1.43
GO:0006310	DNA recombination	109	1.21
GO:0051301	cell division	88	0.98
GO:0007049	cell cycle	83	0.92
GO:0006260	DNA replication	81	0.90
GO:0008610	lipid biosynthetic process	80	0.89
GO:0006807	nitrogen compound metabolic process	68	0.76
GO:0000160	two-component signal transduction system (phosphorelay)	68	0.76
GO:0006811	ion transport	56	0.62
GO:0007059	chromosome segregation	56	0.62
GO:0044237	cellular metabolic process	55	0.61
GO:0006974	response to DNA damage stimulus	54	0.60
GO:0007165	signal transduction	53	0.59
GO:0015074	DNA integration	53	0.59
GO:0006418	tRNA aminoacylation for protein translation	51	0.57
GO:0006313	transposition, DNA-mediated	50	0.56
GO:0006950	response to stress	43	0.48
GO:0007242	intracellular signaling cascade	41	0.46
GO:0045454	cell redox homeostasis	39	0.43
GO:0008033	tRNA processing	39	0.43
GO:0046677	response to antibiotic	38	0.42
GO:0006396	RNA processing	36	0.40
GO:0006352	transcription initiation	36	0.40
GO:0016310	phosphorylation	35	0.39
GO:0007047	cell wall organization	35	0.39
GO:0006468	protein amino acid phosphorylation	34	0.38
GO:0009073	aromatic amino acid family biosynthetic process	33	0.37
GO:0006754	ATP biosynthetic process	33	0.37
GO:0008272	sulfate transport	32	0.36
GO:0009116	nucleoside metabolic process	31	0.35
GO:0006289	nucleotide-excision repair	31	0.35
GO:0006099	tricarboxylic acid cycle	30	0.33
GO:0006164	purine nucleotide biosynthetic process	28	0.31
GO:0006284	base-excision repair	28	0.31
GO:0015031	protein transport	27	0.30
GO:0006541	glutamine metabolic process	27	0.30
GO:0022900	electron transport chain	27	0.30
GO:0009190	cyclic nucleotide biosynthetic process	27	0.30
GO:0009236	cobalamin biosynthetic process	26	0.29
GO:0006725	cellular aromatic compound metabolic process	26	0.29
GO:0009082	branched chain family amino acid biosynthetic process	25	0.28
GO:0009252	peptidoglycan biosynthetic process	25	0.28
GO:0006413	translational initiation	25	0.28
GO:0009273	peptidoglycan-based cell wall biogenesis	24	0.27
GO:0008299	isoprenoid biosynthetic process	24	0.27
GO:0006457	protein folding	24	0.27
GO:0015833	peptide transport	22	0.25
GO:0006779	porphyrin biosynthetic process	22	0.25
GO:0006259	DNA metabolic process	22	0.25
GO:0006520	cellular amino acid metabolic process	22	0.25
GO:0009405	pathogenesis	21	0.23
GO:0006631	fatty acid metabolic process	21	0.23
GO:0015986	ATP synthesis coupled proton transport	21	0.23
GO:0006812	cation transport	20	0.22
GO:0006777	Mo-molybdopterin cofactor biosynthetic process	20	0.22
GO:0008654	phospholipid biosynthetic process	20	0.22
GO:0008360	regulation of cell shape	20	0.22



(a) Biological process ontology.



(b) Molecular function ontology.



(c) Distribution of levels of GO terms predicted.

Figure 4.9: *Pie chart showing GO slim functions of hypothetical proteins for MF and BP and bar chart showing distribution of levels of GO terms predicted.*

Table 4.9: *GO Slim terms (generic GO slim) significantly over-represented in newly predicted GO set compared to complete set of GO terms.*

GO ID	GO name	P-Value
GO:0043170	macromolecule metabolic process	0.000000
GO:0044238	primary metabolic process	0.000000
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	4.42e-12
GO:0008152	metabolic process	7.61e-12
GO:0009987	cellular process	3.90e-11
GO:0019538	protein metabolic process	5.17e-11
GO:0006259	DNA metabolic process	1.17e-10
GO:0003677	DNA binding	1.86e-10
GO:0006350	transcription	4.78e-10
GO:0008233	peptidase activity	8.39e-10
GO:0043283	biopolymer metabolic process	1.94e-08
GO:0044237	cellular metabolic process	1.96e-07
GO:0003676	nucleic acid binding	5.78e-07
GO:0006629	lipid metabolic process	8.71e-06
GO:0016787	hydrolase activity	1.27e-05
GO:0016740	transferase activity	1.45e-05
GO:0050789	regulation of biological process	6.26e-05
GO:0065007	biological regulation	8.98e-05
GO:0005975	carbohydrate metabolic process	0.011421
GO:0006412	translation	0.011646
GO:0009059	macromolecule biosynthetic process	0.011646
GO:0044249	cellular biosynthetic process	0.011646
GO:0007049	cell cycle	0.012909
GO:0004518	nuclease activity	0.014776
GO:0016788	hydrolase activity, acting on ester bonds	0.019536
GO:0003700	transcription factor activity	0.021468

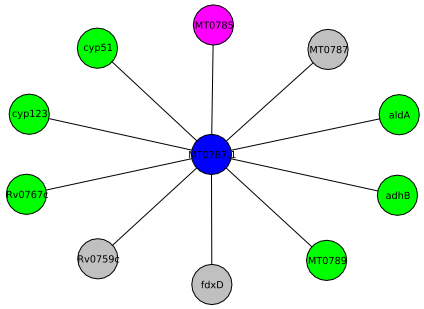
A summary of the main processes in which these proteins are predicted to be involved are shown in table 4.11. These predictions suggest that most of these proteins may be involved in proteolysis, metabolic and oxidation reduction processes. This is in general agreement with the suspicion that these proteins may play an important role in the survival of the MTB pathogen in the host. We can speculate that they may provide the pathogen the ability to switch from one metabolic path to another including aerobic and anaerobic, thus allowing the pathogen to survive within the host in different environments ranging from high oxygen potential in the lungs to micro-aerobic/anaerobic conditions within the tuberculous granuloma.

As an illustration of some example annotations, protein MT0787.1, named Ferredoxin-related protein (UniProt accession Q7D9B4), which is involved in intermediary metabolism and respiration (figure 4.10(a)) but still uncharacterized with respect to GO biological process ontology, is functionally linked to proteins involved in oxidation reduction (GO:0055114) or sharing similarity at a certain level with this GO term. Thus, it is likely that Ferredoxin-related protein MT0787.1 is involved in oxidation reduction. For the protein MT3181 (Q7D648) in figure 4.10(b) belonging to PE/PPE/PGRS family protein,

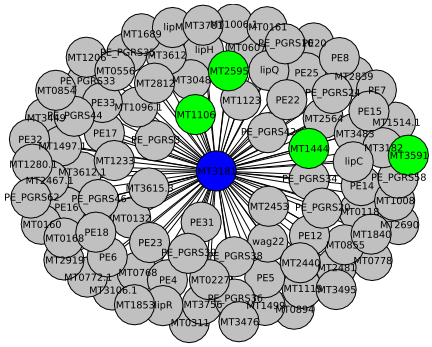
Table 4.10: *Predicted functions with their GO annotation scores for some protein members of the PE/PPE family.*

UniProt-Acc	Gene-Name	GO-ID	GO name	Annotation score
Q8VKA5	MT1008	GO:0006468	protein amino acid phosphorylation	0.58786
		GO:0006508	proteolysis	0.58786
Q7D937	PE7	GO:0008152	metabolic process	0.52573
		GO:0055114	oxidation reduction	0.52573
		GO:0006508	proteolysis	0.37432
Q8VIW6	MT3756	GO:0006508	proteolysis	1.00000
Q50615	PE_PGRS33	GO:0008652	cellular amino acid biosynthetic process	0.35613
		GO:0009082	branched chain family amino acid biosynthetic process	0.35428
		GO:0055085	transmembrane transport	0.29735
		GO:0006810	transport	0.29305
Q8VK65	MT1168	GO:0008152	metabolic process	1.00000
Q8VK71	MT1123	GO:0006508	proteolysis	1.00000
Q8VJ66	MT3247	GO:0006355	regulation of transcription, DNA-dependent	0.33233
		GO:0045449	regulation of transcription	0.33233
		GO:0006725	cellular aromatic compound metabolic process	0.25052
		GO:0055114	oxidation reduction	0.24008
P0A692	ppe30	GO:0008152	metabolic process	0.23468
Q8VKM4	MT0369	GO:0008152	metabolic process	0.83857
		GO:0055114	oxidation reduction	0.67714
Q7D568	PE32	GO:0006508	proteolysis	1.00000
O53416	PE_PGRS20	GO:0006508	proteolysis	1.00000
Q10892	ppe1	GO:0008152	metabolic process	0.83857
		GO:0055114	oxidation reduction	0.67714
Q7DA58	MT0269	GO:0051188	cofactor biosynthetic process	0.26862
		GO:0006541	glutamine metabolic process	0.24554
		GO:0009236	cobalamin biosynthetic process	0.23250
Q7D4P4	MT3987	GO:0051301	cell division	0.31231
		GO:0007059	chromosome segregation	0.31231
		GO:0007049	cell cycle	0.31231
		GO:0045449	regulation of transcription	0.27694
		GO:0006355	regulation of transcription, DNA-dependent	0.27694
Q10778	ppe21	GO:0008152	metabolic process	0.49715
		GO:0006629	lipid metabolic process	0.40134
		GO:0006631	fatty acid metabolic process	0.29640
		GO:0008654	phospholipid biosynthetic process	0.29065
Q7D8N0	MT1233	GO:0006508	proteolysis	1.00000
O86338	PE6	GO:0006508	proteolysis	1.00000
Q7D7Y7	MT1839	GO:0008152	metabolic process	0.28426
Q7D7Y8	MT1838	GO:0055114	oxidation reduction	1.00000
Q7D8Q2	PPE17	GO:0045449	regulation of transcription	0.55388
		GO:0006355	regulation of transcription, DNA-dependent	0.55388
		GO:0006350	transcription	0.33333
P42611	ppe10	GO:0008152	metabolic process	0.24834
Q8VJK7	MT2422	GO:0008152	metabolic process	0.23520
O06246	ppe59	GO:0008152	metabolic process	0.24834
Q7D7Y9	PE18	GO:0055114	oxidation reduction	0.53074
		GO:0006508	proteolysis	0.53074

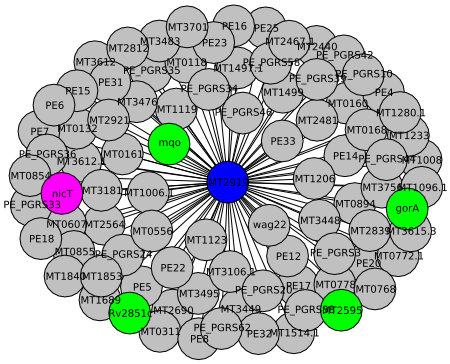
UniProt-Acc	Gene-Name	GO-ID	GO name	GO-score
Q7D5G7	PE31	GO:0006629	lipid metabolic process	0.50899
		GO:0008152	metabolic process	0.48975
		GO:0006508	proteolysis	0.39356
Q8VIY9	MT3615.3	GO:0008152	metabolic process	0.68716
		GO:0006508	proteolysis	0.37432
Q7D8W7	MT1096.1	GO:0006508	proteolysis	1.00000
Q8VIY0	MT3663	GO:0008152	metabolic process	0.24615
Q7D4N3	PE36	GO:0051301	cell division	0.61929
		GO:0007049	cell cycle	0.61929
		GO:0007059	chromosome segregation	0.61929
Q8VJ87	MT3105	GO:0051301	cell division	0.61929
		GO:0007049	cell cycle	0.61929
		GO:0007059	chromosome segregation	0.61929
Q7D7Y6	MT1840	GO:0006508	proteolysis	1.00000
Q7D7X9	PPE29	GO:0008152	metabolic process	0.26232
Q7D974	MT0854.1	GO:0008152	metabolic process	0.57379
		GO:0009058	biosynthetic process	0.55642
		GO:0055114	oxidation reduction	0.48769
Q7D9B9	MT0779	GO:0008152	metabolic process	0.51728
		GO:0006313	transposition, DNA-mediated	0.51728
P31500	ppe46	GO:0008610	lipid biosynthetic process	0.53351
		GO:0006313	transposition, DNA-mediated	0.53351
Q8VIZ0	MT3612.1	GO:0006508	proteolysis	0.38380
		GO:0006313	transposition, DNA-mediated	0.37483
		GO:0008152	metabolic process	0.36535
Q11031	ppe19	GO:0005975	carbohydrate metabolic process	0.45012
		GO:0008610	lipid biosynthetic process	0.43301
		GO:0006013	mannose metabolic process	0.42628
		GO:0006508	proteolysis	0.33941
Q8VJW5	MT1836	GO:0055114	oxidation reduction	1.00000
Q7D8M9	PPE18	GO:0055114	oxidation reduction	1.00000
Q8VIX6	MT3701	GO:0006508	proteolysis	0.47303
		GO:0019538	protein metabolic process	0.47303
		GO:0006289	nucleotide-excision repair	0.37837
Q10637	PE_PGRS24	GO:0006508	proteolysis	1.00000
Q7DAC9	PE4	GO:0006508	proteolysis	1.00000
Q10540	ppe13	GO:0008152	metabolic process	0.24834
Q8VKC5	MT0894	GO:0008152	metabolic process	0.52573
		GO:0055114	oxidation reduction	0.52573
		GO:0006508	proteolysis	0.37432
Q7D7S8	MT1969	GO:0008152	metabolic process	0.28769



(a) Direct interacting partners of protein 'MT0787.1' (Intermediary metabolism and respiration protein) predicted to be involved in oxidation reduction.



(b) Direct interacting partners of protein 'MT3181' (PE/PPE protein) predicted to be involved in Proteolysis.



(c) Direct interacting partners of protein 'M2919' (PE/PPE protein) predicted to be involved metabolic process.

Figure 4.10: *Illustration of Annotation inference. Protein target is in blue at the center, proteins in green are those sharing GO similarity at a certain level and those in pink are those with no similar term with the one predicted for the protein target. Proteins in gray are uncharacterized with respect to BP ontology.*

specifically to the PE family proteins, all its annotated interacting partners are involved in proteolysis process (GO:0006508) or share similarity at a certain level with GO:0006508. In this case, the most probable biological process in which the protein MT3181 may be involved is proteolysis. Similarly, the protein MT2919 (Q8VJC0) in figure 4.10(c) belongs to PE/PPE/PGRS family protein, more precisely the PE-PGRS family protein, and most of its annotated interacting partners are involved in metabolic process (GO:0008152) or

Table 4.11: *Different processes in which PE-PPE proteins are mostly involved.*

GO-ID Process	GO name	Number of Proteins	Weight
GO:0006508	proteolysis	68	21.3
GO:0008152	metabolic process	51	16.0
GO:0055114	oxidation reduction	24	7.5
GO:0006355	regulation of transcription, DNA-dependent	9	2.82
GO:0045449	regulation of transcription	9	2.82
GO:0008652	cellular amino acid biosynthetic process	8	2.51
GO:0006810	transport	8	2.51
GO:0005975	carbohydrate metabolic process	7	2.19
GO:0006629	lipid metabolic process	5	1.57
GO:0009058	biosynthetic process	5	1.57
GO:0055085	transmembrane transport	5	1.57
GO:0051301	cell division	5	1.57
GO:0007059	chromosome segregation	5	1.57
GO:0007049	cell cycle	5	1.57
GO:0007242	intracellular signaling cascade	4	1.25
GO:0008610	lipid biosynthetic process	4	1.25
GO:0006313	transposition, DNA-mediated	4	1.25

share similarity at a certain level with GO:0008152. Thus, its predicted GO biological process is metabolic process.

4.6 Summary

In this chapter we perform a functional analysis of the MTB network through protein function prediction using a local neighborhood (level-1 and level-2) based approach. The key idea driving the choice of the approach used to predict protein function is the underlying biological principle of the MTB functional network structure, allowing the system to be stable and robust, functioning in a reliable way. We predicted functions of a large number of uncharacterized proteins in MTB CDC1551 using GO biological process and molecular function terms, using the GO-universal similarity metric to compare these GO terms. This has yielded additional information about the biology of MTB CDC1551 with 3804, and 3698 proteins with predicted functions using the GO biological process and molecular function ontologies, representing approximately 90.7% and 88.2% of the MTB proteome, respectively. Further analysis of these predictions has revealed that most of MTB proteins are involved in metabolic and oxidation reduction processes.

Chapter 5

Structural Analysis of the MTB Proteome Networks

The biological analysis of organisms has evolved from a one-by-one gene approach to the whole genome focus, providing the opportunity to look at genes within their context in a cell and achieve a global view. For the organism under study, namely *Mycobacterium tuberculosis* (MTB), the contribution of knowledge discovered from both primary data, such as genomic sequences and functional data from high-throughput experiments has led to the construction of a protein-protein functional network. Such a network allows the organism's proteome to be handled in a global fashion in order to unravel the underlying principles of its biological properties for the purpose of building a predictive disease model and identifying novel therapeutic drug targets for disease. The network obtained represents all protein pair interactions derived from different biological data sources. Thus, after performing the functional analysis of the network (chapter 4), the next step is to explore (1) the interplay between each protein pair in the network and their contribution to disease, and (2) how they reliably function for the survival and fitness of the organism on the basis of the network topology. This may help identify proteins which enable the organism to efficiently carry out its high goal in the host.

In this chapter, we perform extensive computations to detect the key principles driving the biological organization of the organism, and decorticate the system to identify proteins that are potentially indispensable for the survival and viability of the organism, referred to as *essential proteins*, and those which contribute to the fitness of the organism, referred to

as *supplementary proteins*. This categorization can provide clues toward finding effective drug targets. Essential proteins are ultimate candidates [234], and could possibly lead to new anti-tuberculosis compounds with novel mechanisms of action. Of particular interest in these rationale target-based strategies, are proteins identified by previous research projects to be unique to *Mycobacterium tuberculosis* or to pathogens including protein members of the *Pro-Glu* and *Pro-Pro-Glu* (*PE/PPE*) and *Polymorphic GC-Rich Sequence* (*PE-PGRS*) families. They have been suspected to play a role in the virulence or immunogenicity of *tuberculosis* [235, 236] by altering the way the host responds to the infection. They may have a particular role in helping the organism evade the host immune response and in latency, and are thus likely to be important for the specific lifestyle of the organism. Along with them, a significant proportion of proteins labelled hypothetical and uncharacterized in protein sequence databases are unique to these pathogens.

In the following section, we describe several measures of network centrality that are later applied to the MTB functional network to identify structurally important proteins in the network. Thereafter, we perform biological analysis of these proteins to determine which of those may be essential for the organism based on the topology of the network. Finally, we map suspected virulence genes and hypothetical proteins of interest onto the functional network to identify the neighborhood of those proteins effectively playing a key role within the bacterial infection and disease-causing process.

5.1 Topological Network Centrality Measures

In order to understand the biological organization of the organism from its protein functional network and use this as a means to develop appropriate treatment strategies for the disease, complete knowledge of the network structure and the contribution of each protein to the system's biological processes are required. This section describes network measures that are used to numerically characterize the importance of proteins in the system, and their contribution to the functioning of the system, thus assessing the topological significance of these proteins within the network and quantifying the structural properties of the functional network produced. These measures include degree or connectivity, betweenness, closeness and eigenvector centrality metrics. These measures are used to reveal proteins

which are potentially crucial to the functioning of the system, thus contributing to the survival of the organism.

Throughout this section, we denote by $G(\mathcal{N}, \mathcal{L})$ the MTB functional network, with $\mathcal{N}$ the set of interacting proteins and $\mathcal{L}$ the set of functional interactions or connections between proteins, represented by the adjacency matrix $\mathcal{A}$, an $n \times n$ symmetric matrix, where $n = |\mathcal{N}|$ is the number of proteins in the network and whose components a_{pq} are defined as follows:

$$a_{pq} = \begin{cases} 1 & \text{if the protein } p \text{ is functionally linked to the protein } q, \\ 0 & \text{otherwise.} \end{cases}$$

Here, proteins in $\mathcal{N}$ are numbered from 1 to n, and a protein p is represented by its position number denoted by p. The adjacency matrix $\mathcal{A}$ is symmetric since if the protein p is functionally linked to the protein q, then clearly the protein q is also functionally linked to the protein p. Note that a given protein p is not functionally linked or connected to itself, *i.e.*, $a_{pp} = 0$.
$\pi(p, q)$ denotes the distance between proteins p and q or the length of the shortest path from a protein p to a protein q, *i.e.*, the number of links in the shortest path between p and q for an unweighted graph; the shortest path between proteins being the path with the minimum number of edges connecting these proteins. If no path exists between proteins p and q then $\pi(p, q) = \infty$.

5.1.1 Degree and Betweenness Centrality Metrics

The degree or connectivity of a protein p is the number of links connected to it, *i.e.*, the number of its interacting neighbors [237] given by

$$deg(p) = \sum_{q \in \mathcal{N}} \delta(p, q) \tag{5.1}$$

where

$$\delta(p, q) = \begin{cases} 1 & \text{if the protein } q \text{ is functionally linked to the protein } p \\ 0 & \text{otherwise.} \end{cases}$$

In terms of the adjacency matrix $\mathcal{A}$, the degree of a protein p is simply the sum of components in the row or the column corresponding to the protein p, given by

$$\sum_{q=1}^{n} a_{pq} = deg\,(p) = \sum_{q=1}^{n} a_{qp}. \tag{5.2}$$

In fact, the degree or connectivity of a protein provides an indicator of its influence on the biological processes occurring in the organism meaning that a protein with more functional connections tends to contribute to several processes, and may thus be a key protein in the functioning of the system.

The betweenness centrality of a protein p in a functional network is a metric that expresses the influence of p relative to other proteins within the network. It is based on the proportion of shortest paths between other proteins passing through the protein target [238] and shows the importance of a protein for the transmission of information between other proteins in the network. This metric provides an indication of the number of pair-wise proteins connected indirectly by the protein target through their direct functional connections. The betweenness, $B\,(p)$, of a protein p is given by

$$B\,(p) = \sum_{(s,t)\in\mathcal{N}_p} \frac{\sigma_{st}\,(p)}{\sigma_{st}} \tag{5.3}$$

where $\sigma_{st}\,(p)$ is the number of shortest paths from protein s to protein t passing through p, σ_{st} the number of shortest paths from s to t in the functional network, and $\mathcal{N}_p = \{(s,t) \in \mathcal{N} \times \mathcal{N} : s \neq p \neq t \text{ and } s \neq t\}$.
The normalized betweenness of a protein p, lying between 0 and 1 is given by

$$B\,(p) = \frac{1}{(n-1)\,(n-2)} \sum_{(s,t)\in\mathcal{N}_p} \frac{\sigma_{st}\,(p)}{\sigma_{st}} \tag{5.4}$$

Thus, proteins with high betweenness are expected to ensure the connectivity between proteins in the functional network and are able to bridge or disconnect connected components. As the MTB functional network generated has a scale free property, such proteins are hubs, referring to proteins that are highly connected and serve to hold together a large number

of proteins with low degree, thus integrating all proteins in a given connected component into a unified complex system. These proteins are of utmost importance for the integrity and the robustness of the system and are responsible for the small world property since connections between proteins are relatively short via these hubs.

5.1.2 Closeness and Confidence Measures of a Protein

The status, $S(p)$, of a protein p in a connected network is the average distance to all other proteins, *i.e*, the ratio of the sum of $\pi(p,q)$ for all proteins q in the network to the total possible number of such paths, which is $n(n-1)$. It is given by

$$S(p) = \frac{1}{n(n-1)} \sum_{q \in \mathcal{N}} \pi(p,q) \tag{5.5}$$

The closeness measure, $\mathcal{C}_s(p)$, of a protein p is the inverse [237] of its status and reflects the ability of the protein to access information via other proteins and to propagate information through the network. As the MTB functional network is not completely connected, this closeness measure is calculated for each connected part separately and normalized to $(n_c - 1)/(|\mathcal{L}_c| - 1)$ [239], where n_c is the number of nodes in the connected part of the graph containing the node and $|\mathcal{L}_c|$ its size, *i.e*., the number of functional links in the connected component. This is to make the scale uniform for comparison. Thus, the closeness measure of a protein p is given by

$$\mathcal{C}_s(p) = \frac{|\mathcal{L}_c| - 1}{(n_c - 1) \times S_r(p)} \tag{5.6}$$

where $S_r(p)$ is the status of p relative to its connected component.
The closeness measure is high for a protein that is central since it has a shorter distance on average to other proteins. We define the center of gravity $\mathcal{G}_c$ of the network as the set of proteins that maximize the closeness measure to any other protein in the network, given by

$$\mathcal{G}_c = \left\{ p \in \mathcal{N} : \mathcal{C}_s(p) = \max_{q \in \mathcal{N}} \mathcal{C}_s(q) \right\} \tag{5.7}$$

The eccentricity, $E(p)$, of a protein p in a given connected graph is the maximum length of shortest paths from protein p to all other proteins in the network, *i.e.*,

$$E(p) = \max\{\pi(p,q) : q \in \mathcal{N}\} \tag{5.8}$$

In the context of the MTB functional network, the eccentricity $E(p)$ of a protein p is computed according to its connected component and we consider the inverse of the eccentricity obtained, and normalize it, as done previously. The measure is referred to as the confidence $\mathcal{C}_e(p)$ of protein p, expressing its capability to quickly communicate with other proteins in the network, and given by:

$$\mathcal{C}_e(p) = \frac{|\mathcal{L}_c| - 1}{(n_c - 1) \times E_r(p)} \tag{5.9}$$

where $E_r(p)$ is the eccentricity of p relative to its connected component.
The higher the confidence of a protein in the functional network, the quicker it communicates with other proteins in the network. We define the reference center $\mathcal{R}_c$ of the network as the set of proteins that maximize the confidence of any other protein in the network, given by

$$\mathcal{R}_c = \left\{ p \in \mathcal{N} : \mathcal{C}_e(p) = \max_{q \in \mathcal{N}} \mathcal{C}_e(q) \right\} \tag{5.10}$$

5.1.3 Eigenvector Centrality Metric

The degree or connectivity metric provides a simple number of functional connections without weighting them. The eigenvector metric considers the importance or weight of these functional connections [237]. In fact, functional connections are not equally important and functional connections to more influential proteins will impact more on the contribution of the protein than functional connections to less influential proteins. Thus, the eigenvector centrality metric assigns a relative weight to all proteins in the network based on the fact that functional connections to proteins of high weight contribute more to the weight of the protein target.

Let c_p be the numerical value representing the contribution of the protein p to the functioning of the system. c_p is then proportional to the contributions of its neighbors to the system. This means that

$$\sum_{q=1}^{n} a_{pq} c_q = \lambda c_p \tag{5.11}$$

where λ is constant for every protein p in the functional network. In matrix form, this can be written as follows:

$$\mathcal{A}c = \lambda c \tag{5.12}$$

where $c = (c_1, \ldots, c_n)^T$, the transpose of the vector $(c_1, \ldots, c_n)$, which defines a vector of contributions of each protein. The vector c is an eigenvector of the adjacency matrix $\mathcal{A}$ associated with eigenvalue λ. It is known that λ is the largest eigenvalue of the adjacency matrix and c is its non-negative corresponding eigenvector [237, 240].

In this metric, the contribution of a given protein to the functioning of the system depends not only on the number of its interacting neighbors but also on the quality of these neighbors. Proteins with a high number of functional interactions are important, but a protein with a small number of high-quality functional connections may contribute more to the survival of the organism than one with a large number of low-quality functional connections.

5.2 Topological Analysis of the MTB Functional Network

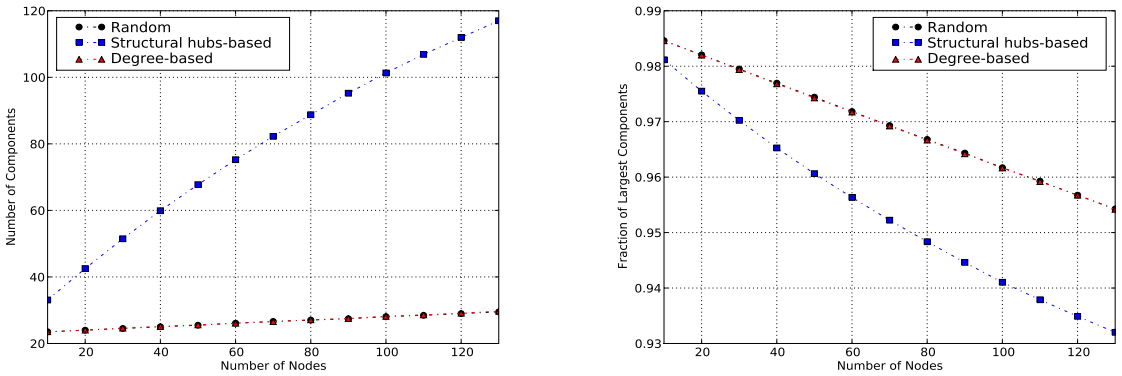
The robustness of the system is observed through its stability, expressed by its ability to remain non vulnerable under changing environmental conditions or stressful perturbations due to a protein knock-out or attack. Topologically, this can be seen as the potential connectivity of the network under a protein disruption. The disruption can be random, in which case it is called error or random attack, or targeted, in which case the protein

assumed to be essential for the system functioning is disrupted, referred to as targeted attack.

5.2.1 Assessing High-Degree Proteins

We have previously seen that the MTB functional network exhibits a 'scale-free' property. As such, it is expected to be vulnerable against targeted attack and robust against random attack. In order to assess the topological essentiality of MTB proteins, we classify them in two categories, namely proteins with a high degree referred to as degree-based hubs and those able to disconnect the functional network, known as structural hubs. A protein is considered to be a degree based hub if its degree is above the average degree of proteins in the MTB functional network, which is 28. We first observe the changes in the number of connected components and in the number of proteins in the largest connected component by repeatedly (1) knocking out randomly selected proteins referred to as random attacks, (2) disrupting the highest degree proteins, referred to as degree-based hub attack, and (3) removing proteins able to disconnect the network, referred to as structural hub attack. To simulate an attack, a given number of proteins is chosen for each category and the process is repeated 1000 times by randomly choosing proteins and computing the average number of the resulting components and the number of proteins in the largest component. Results are shown in figure 5.1 and indicate that the MTB functional network is vulnerable to targeted structural hub attacks.

Indeed, the more structural hubs that are removed the higher the number of connected components. This means that the more structural hubs that are removed, the more the network is disintegrated, whereas the disruption of randomly selected proteins, or of degree-based hubs, does not perturb the general structure of the network. This means that structural hub proteins play an essential role in the network integrity. Therefore, knocking out these proteins may disturb the functioning of the system and negatively impact on the ability of this pathogenic bacterium to carry out its higher goal in the host.
We have also analyzed the network connectivity by observing the size of the largest connected component. Figure 5.1(b) shows that the size of the largest component rapidly decreases when structural hubs are disrupted. This indicates that the network is disintegrated into several small connected components, thus showing the role played by the



(a) Variations in the number of connected components.

(b) Variations in the number of proteins in the largest connected component.

Figure 5.1: *Assessing network vulnerability under random and targeted attacks.*

structural hubs in maintaining the network connectivity.

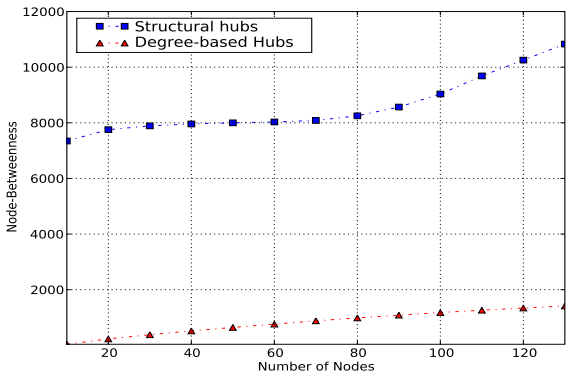
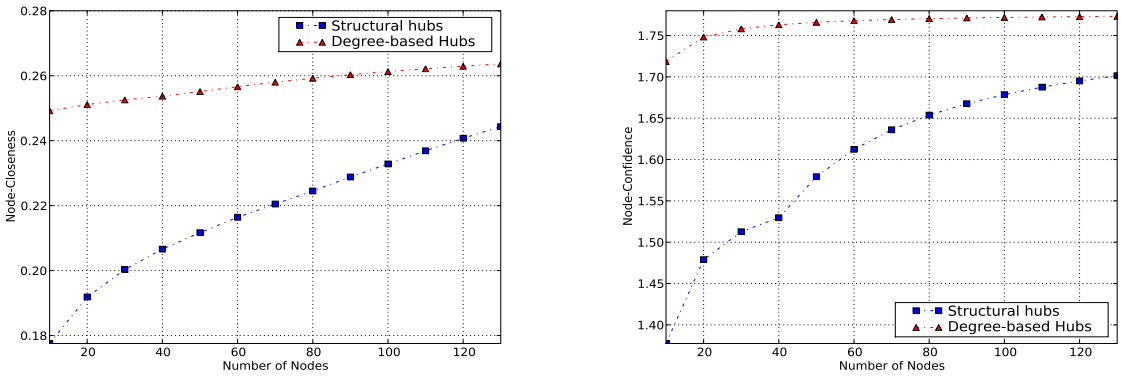


Figure 5.2: *Analyzing the variations in the betweenness metric in terms of protein category.*

5.2.2 Assessing Central Proteins

The betweenness metric represents a significant indicator of network essentiality [161]. Proteins with high betweenness are essential to the functioning of the system, serving as bridges for communication between several other proteins in the network. A protein with high confidence or closeness will be more important because it has a smaller path length to reach all other proteins in the network, allowing the system to quickly exhibit appropriate behaviour in case of a given perturbation in the system. Figures 5.2 and 5.3 show the



(a) Variations in the closeness measure.

(b) Variations in the confidence metric.

Figure 5.3: *Assessing the variations in closeness and confidence centrality measures in terms of protein category.*

functional importance of proteins obtained by ordering proteins by betweenness, closeness and confidence measures of hubs, and observing the cumulative proportion for every 10 proteins. These results reveal that proteins with high degrees and located in the center of the network may reach all the nodes in a given connected component with fewer steps compared to the structural hubs. These results combined with those in figures 5.3 suggest that a protein, which is structural hub and has a high degree, is important for the survival of the bacterial pathogen. These proteins are considered to be potential drug targets and can be used to enhance the discovery process of new antibiotics with novel mechanisms of action to treat the disease.

5.3 Important Proteins in the MTB Functional Network

In this section, we are investigating the biological significance of proteins found to be structurally important in the functional network. Specifically, we are looking at the functions that are carried out by proteins found in the center of gravity $\mathcal{G}_c$ with high betweenness and connected to some influential proteins at certain levels, *i.e.*, proteins with eigenvector centrality greater than 10^{-5}. We are also interested in the biological processes in which they are involved, as well as in the functional class to which they belong. This enables

the identification of proteins that are potentially essential for the survival of the bacterial pathogen. These proteins are considered to be potential drug targets, as they correspond to bottlenecks in the MTB functional network and are therefore expected to be key components of the organism's cellular processes. Bottleneck proteins are proteins responsible for several indirect functional connections between other proteins in the functional network. Note that the confidence metric is not used since there is no discerning difference between eccentricity distribution of hubs and non hubs. Thus, this metric could not be used to distinguish essential proteins from non-essential proteins.

As the average shortest path length is 3.678, a protein in the functional network is said to belong to the gravity center if its closeness metric, as defined in equation (5.6) is greater than 1/3.678, which is approximately 0.27189. In the case of the betweenness measure, a protein with betweenness above the total number of shortest paths expected to pass through the protein in the functional network is of interest, and this number is about 15212.21. Thus, we identified a set of 881 proteins, which constitutes a set of important proteins and thus potential drug targets within the bacterial pathogen.

Furthermore, we use function prediction of uncharacterized proteins performed in section 4.4 of chapter 4, leaving out proteins which remained uncharacterized to refine the results. Before running the prediction annotation algorithm, 292 proteins out of the 881 were uncharacterized with respect to GO biological process terms. After deploying the prediction annotation algorithm, functions of 245 proteins were predicted and 47 remain uncharacterized. The resulting dataset consists of a target set of 834 proteins and the analysis performed on this set is shown in table 5.1, and indicates that of the 834 target proteins, 347 are involved in metabolic processes, 226 in oxidation reduction processes, and a significant number of target proteins are involved in regulation, transport, proteolysis, etc. Note that a given protein may be involved in multiple processes.

Further functional analysis was performed using the small group of high level functional classes assigned to all the proteins. The distribution of these potential drug targets per functional class is shown in figure 5.4. These results indicate that most of the protein candidate drug targets are involved in intermediary metabolism, followed by a significant proportion of proteins in the unknown class and those belonging to the cell wall and cell process functional classes. We also identify within the candidate drug target list, proteins

Table 5.1: *Different processes in which MTB potential drug targets are mostly involved.*

GO-ID Process	GO name	Number of Proteins in	Weight
GO:0008152	metabolic process	347	15.78
GO:0055114	oxidation reduction	226	10.28
GO:0006355	regulation of transcription, DNA-dependent	92	4.18
GO:0045449	regulation of transcription	86	3.91
GO:0006350	transcription	85	3.87
GO:0006810	transport	71	3.23
GO:0009058	biosynthetic process	55	2.50
GO:0006412	translation	47	2.14
GO:0006508	proteolysis	43	1.96
GO:0006629	lipid metabolic process	36	1.64
GO:0005975	carbohydrate metabolic process	34	1.55
GO:0055085	transmembrane transport	33	1.50
GO:0008652	cellular amino acid biosynthetic process	33	1.50
GO:0006281	DNA repair	30	1.36
GO:0051301	cell division	29	1.32
GO:0007049	cell cycle	27	1.23
GO:0000160	two-component signal transduction system (phosphorelay)	24	1.09
GO:0044237	cellular metabolic process	23	1.05
GO:0006974	response to DNA damage stimulus	23	1.05
GO:0007059	chromosome segregation	20	0.91
GO:0006418	tRNA aminoacylation for protein translation	18	0.82
GO:0006260	DNA replication	18	0.82
GO:0008610	lipid biosynthetic process	17	0.77
GO:0006811	ion transport	16	0.73
GO:0046677	response to antibiotic	15	0.68
GO:0006310	DNA recombination	14	0.64
GO:0006807	nitrogen compound metabolic process	13	0.59
GO:0006950	response to stress	12	0.55
GO:0006096	glycolysis	10	0.45
GO:0007047	cell wall organization	10	0.45
GO:0016310	phosphorylation	10	0.45
GO:0006164	purine nucleotide biosynthetic process	10	0.45
GO:0006541	glutamine metabolic process	10	0.45
GO:0006099	tricarboxylic acid cycle	9	0.41
GO:0007242	intracellular signaling cascade	9	0.41
GO:0045454	cell redox homeostasis	9	0.41
GO:0006221	pyrimidine nucleotide biosynthetic process	9	0.41
GO:0015833	peptide transport	9	0.41
GO:0015074	DNA integration	9	0.41
GO:0007165	signal transduction	9	0.41
GO:0006396	RNA processing	8	0.36
GO:0006313	transposition, DNA-mediated	8	0.36
GO:0006468	protein amino acid phosphorylation	8	0.36
GO:0006520	cellular amino acid metabolic process	8	0.36
GO:0009432	SOS response	8	0.36
GO:0006730	one-carbon metabolic process	8	0.36
GO:0006457	protein folding	8	0.36
GO:0009116	nucleoside metabolic process	7	0.32
GO:0009082	branched chain family amino acid biosynthetic process	7	0.32
GO:0009097	isoleucine biosynthetic process	7	0.32
GO:0009252	peptidoglycan biosynthetic process	7	0.32
GO:0006633	fatty acid biosynthetic process	7	0.32
GO:0006754	ATP biosynthetic process	7	0.32
GO:0006352	transcription initiation	7	0.32
GO:0008272	sulfate transport	7	0.32
GO:0006779	porphyrin biosynthetic process	7	0.32
GO:0006413	translational initiation	6	0.27
GO:0009405	pathogenesis	6	0.27
GO:0006526	arginine biosynthetic process	6	0.27
GO:0008654	phospholipid biosynthetic process	6	0.27
GO:0015986	ATP synthesis coupled proton transport	6	0.27
GO:0008360	regulation of cell shape	6	0.27
GO:0006289	nucleotide-excision repair	6	0.27
GO:0009190	cyclic nucleotide biosynthetic process	6	0.27
GO:0022900	electron transport chain	6	0.27

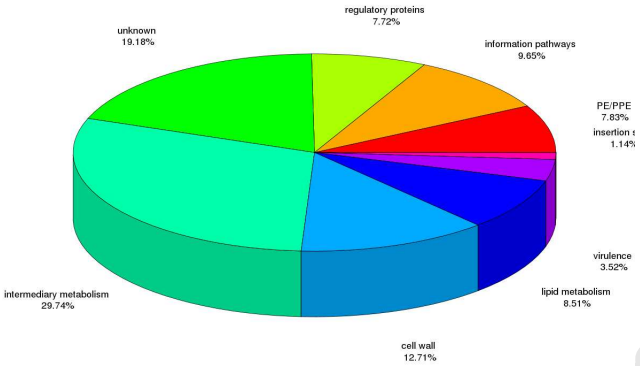


Figure 5.4: *Distribution of candidate drug targets per functional class.*

Table 5.2: *Repartition per class of potential drug target proteins, considering those which are central and those considered to be more influential.*

	Functional Class	Proteins	Drug targets	Central Targets	Influential Targets
1	virulence, detoxification, adaptation	176	31	2	1
2	lipid metabolism	230	75	35	28
3	information pathways	245	85	21	-
4	cell wall and cell processes	618	112	52	5
5	insertion seqs and phages	82	10	-	-
6	PE/PPE	147	69	2	-
7	intermediary metabolism and respiration	884	262	93	70
8	unknown	1637	169	24	10
9	regulatory proteins	176	68	12	-
	Total	4195	881	241	114

which are either more central or more influential in the system and classified them per functional class. Results are shown in figure 5.5 and figure 5.6, and in table 5.2. These results show that most of the potential drug targets that are central to the functioning of the system, ensuring quick communication between proteins in the system, are involved in intermediary metabolism and respiration, cell wall and cell processes, and lipid metabolism. Those involved in intermediary metabolism and respiration, as well as lipid metabolism, are connected to proteins participating in several processes, thus playing key roles in the system.

We used the Fishers Exact Test to find over-represented functions in sets of proteins

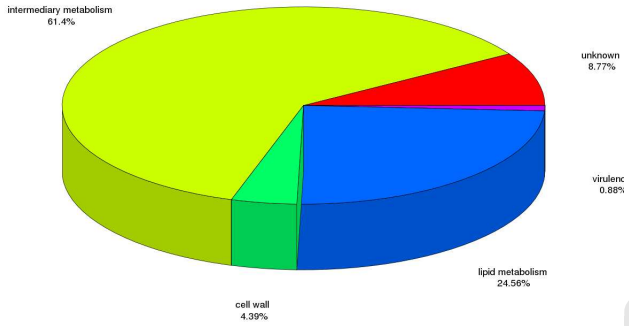


Figure 5.5: *Distribution of more influential drug targets per functional class.*

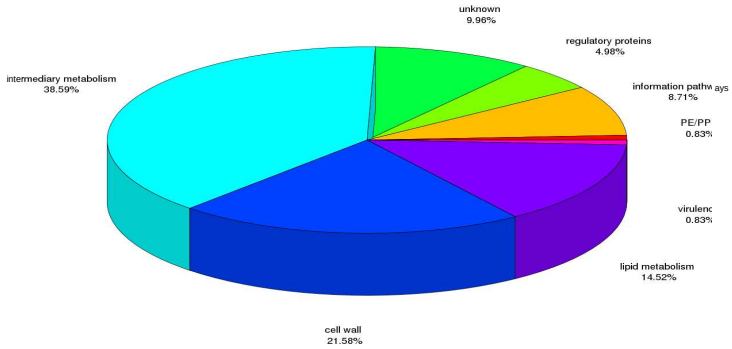


Figure 5.6: *Distribution of more central drug targets per functional class.*

with different network properties. Table 5.3 shows that the hub, high degree (50-99) and high betweenness proteins, as well as the predicted drug target list are significantly over-represented by PE/PPE proteins. Lipid metabolism proteins also feature highly in most of these lists, and regulatory proteins in some of these. One would expect regulatory proteins to be reasonably well connected as they are likely to have an effect on multiple genes. Interestingly, the high closeness measure proteins tend to be from the unknown class. The drug target list also contains 31 proteins belonging to the virulence, detoxification and adaptation functional class.

Table 5.3: *Summary of over-representation analysis of functional classes for different protein sets based on network properties.*

Protein set	Over-represented function	P-value	Adjusted P-value
Hubs	PE/PPE	2.10576e-05	1.89518e-04
Degree >= 100	lipid metabolism	4.37537e-12	1.96891e-11
Degree $50-99$	intermediary metabolism and respiration	1.06668e-25	9.60013e-25
	lipid metabolism	2.33426e-08	7.00278e-08
	information pathways	0.0209259	0.0470832
	regulatory proteins	1.91556e-52	8.62003e-52
	PE/PPE	8.196e-115	7.3764e-114
Degree $10-49$	lipid metabolism	0.00358115	0.0080576
	intermediary metabolism and respiration	1.33874e-58	1.20487e-57
	information pathways	2.6561e-10	7.96829e-10
	virulence, detoxification, adaptation	4.90211e-11	2.20595e-10
Degree < 10	unknown	4.98171e-180	4.48354e-179
	cell wall and cell processes	4.47945e-04	1.78646e-03
	insertion seqs and phages	5.95487e-04	1.78646e-03
Betweenness > 15 000	lipid metabolism	2.03723e-04	3.66702e-04
	intermediary metabolism and respiration	5.99428e-08	1.79828e-07
	information pathways	1.54837e-06	3.48383e-06
	regulatory proteins	3.51658e-08	1.58246e-07
	PE/PPE	4.48875e-11	4.03987e-10
Closeness > 0.5	unknown	2.58864e-14	2.32978e-13
Eigenvector > 0.08	lipid metabolism	1.5511e-12	6.97994e-12
	intermediary metabolism and respiration	2.85447e-31	2.56902e-30
Drug Target	lipid metabolism	2.68651e-05	4.83571e-5
	intermediary metabolism and respiration	2.12524e-11	9.56358e-11
	information pathways	4.13904e-07	9.31285e-07
	regulatory proteins	6.36758e-08	1.91027e-07
	PE/PPE	1.53973e-12	1.38576e-11

Table 5.4: *Summary of network properties of protein sets from the total proteome in the network, those required for normal growth and those required for survival during infection.*

Metric	Total	Growth	Survival
Average Eigenvector	0.003403	0.004342	0.003486
Average Betweenness	10792.87	16108.18	11487.32
Average Closeness	0.28629	0.298827	0.287806
Average Degree	28.082	36.95911	33.17778
% Hubs	4.859768	0.851789	3.888889

Sassetti and colleagues [241, 242] published two lists of genes from MTB H37Rv that have been shown to be involved in either normal growth or for survival during infection. These genes were mapped to CDC1551 identifiers using the orthologues file from the EBI Integr8 project and the network properties for these genes are summarised in table 5.4.

The set of genes required for normal mycobacterial growth tend to have higher average Eigenvector, betweenness, closeness and degree values than the overall proteome. For those required for infection, these values are generally higher than the total average, but not as high as for the growth set. Of the 881 drug target proteins, 197 are on the list of proteins required for growth, and 38 are on the list of proteins required for survival during infection. This enhances their suitability as drug targets, since they have been shown experimentally to be required by the organism.

5.4 Summary

In this chapter, we perform a structural analysis of the MTB CDC1551 functional network, which has enabled potential drug target identification within the bacterial pathogen. This analysis shows that 881 proteins out of 4136 found in the network were predicted to be essential for the functioning of the system and may therefore contribute to the survival of the bacterial pathogen in the host. After removing proteins which remained uncharacterized after applying the protein function prediction algorithm, 834 potentially essential proteins were left. These proteins can form a target list for new drug development and for screening available drugs for the treatment of tuberculosis. However, they would need to go through a number of other screens first in terms of identifying related proteins in the organism itself and the host, etc. to determine their suitability as drug targets.

Chapter 6

General Conclusions

Tuberculosis (TB) remains a public health challenge today, claiming millions of lives and new cases every year. The existing antibiotics, together with enhanced provision of services in recent years, such as Direct Observed Treatment (DOT), are of immense value in controlling the disease. However, these drugs have several shortcomings, the most important being the emergence of drug resistance, making even the front-line drugs ineffective. In addition, the deadly interaction between TB and HIV/AIDS is threatening to compromise gains in TB control, leading to further challenges for anti-tubercular drug discovery. This clearly indicates that the goal of eradicating TB in the coming years depends on the development of new diagnostics, drugs and vaccines.

The identification of drug targets has traditionally been achieved using the complete knowledge of individual proteins and their well characterized functions. Unfortunately, this strategy has failed to deliver a sufficient molecular diversity of drugs in order to overcome this public health challenge. In this work, we have integrated biological data from multiple sources into a single network of protein functional interactions for the analysis of *Mycobacterium tuberculosis* (MTB) uncharacterized genes. This provides a systems view of the whole organism for the identification of potential drug targets within the bacterial pathogen. In order to integrate the data and use it for function prediction, we developed new scoring methods for data from microarray experiments and from protein sequence and signature data, and also developed a novel GO semantic similarity metric. All the data and results have been stored in a MySQL database and made accessible via a web interface, which includes functional link partners and scores for each protein, as well as database

links to KEGG, GO, etc.

We identified potential drug targets through functional and structural analyses of the MTB functional network produced. Functional analysis was performed through function prediction, where possible, of uncharacterized proteins, many of which are specific to the mycobacteria and suspected to play a role in the virulence of MTB. The structure of the network has been analyzed using network centrality measures to identify proteins which maintain the MTB system's stability and robustness, thus helping the bacterial pathogen to survive and achieve its high goal within the host. This has yielded a new predicted MTB biology in which 3804 proteins out of 4195 have predicted annotations with GO biological process terms, of which 834 were identified as potential drug targets within MTB. 3698 proteins have predicted annotations with GO molecular function terms.

This approach may answer critical questions that are often difficult to address experimentally and provide a rational drug target identification mechanism at the molecular level for the disease. It also helps us to better understand the biology of the organism as a whole system and may serve to narrow down the scale of further high throughput target screening. Furthermore, this may contribute to the process of developing new antibiotics with novel mechanisms of action for better treatment of the disease by saving time and reducing the cost.

There is room for further research to complement and extend the work presented in this thesis. These include the following aspects:

1. The study of the system's behavior under changing environmental conditions. This essentially consists of identifying individual components of the system, which can be used to predict the probable response of the system to perturbation. For example, the likely response of the system to a change in a specific gene or protein. This is useful if one wants to measure the impact of drugs on the behavior of the system. The use of available expression data by analyzing the expression of the gene target and the identification of co-expressed genes can provide a solution to this issue. This analysis is an interesting direction for future research.

2. In this thesis, we analyze the virulence genes within MTB, strain CDC1551 and identify

proteins which are potential drug targets. As the disease is a balance between bacterial pathogen virulence and host resistance, knocking out a given protein within the parasite may adversely impact the host system. This means that for a drug target to be effective it must take into account the host system. There is therefore a need to consider the host system in order to produce a comprehensive map of protein interactions between pathogen and the human host. The exploration of this map can yield drug targets which also consider the host system to prevent potential adverse reactions in the host. This issue will be addressed by a future research study.

3. We used the "LAMP" system (Linux, Apache, MySQL and PHP/Python) for implementing our system and constructing an MTB strain CDC1551 protein interaction database with interactive visualization provided via a web interface. This may be extended to include other strains, such as H37Rv, H37Ra, as well as other organisms, which would enable a systems analysis of these organisms, as well as comparative genomic analysis. This can help in the process of validating experiments and also serve to reduce the scale of further high throughput screening for these organisms too. The design and implementation of such a scheme forms another direction for future work.

Bibliography

[1] H. Zheng, L. Lu, B. Wang, S. Pu, X. Zhang, G. Zhu, W. Shi, L. Zhang, H. Wang, S. Wang, G. Zhao, and Y. Zhang. *Genetic Basis of Virulence Attenuation Revealed by Comparative Genomic Analysis of Mycobacterium tuberculosis Strain H37Ra versus H37Rv.* PLoS One, 3(6):e2375-e2375, 2008.

[2] R. Brosch, V. Gordon, K. Eiglmeier, T. Garnier, F. Tekala, E. Yeramian, and S. T. Cole. *Genomics, Biology and Evolution of the Mycobacterium tuberculosis Complex.* In Molecular Genetics of Mycobacteria, 19-36, 2000.

[3] L. Salaun, S. Ayraud, and N. J. Saunders. *Phase variation mediated niche adaptation during prolonged experimental murine infection with Helicobacter pylori.* Microbiology, 151:917-923, 2005.

[4] P. R. Marri, J. P. Bannantine, and G. B. Golding. *Comparative genomics of metabolic pathways in Mycobacterium species: gene duplication, gene delay and lateral gene transfer.* FEMS Microbiol, 30:906-925, 2006.

[5] M. Y. Galperin and E. V. Koonin. *Searching for drug targets in microbial genomes.* Pharmaceutical biotechnology, 10:571-578, 1999.

[6] S. M. Asif, A. Asad, A. Faizan, M. S. Anjali, A. Arvind, K. Neelesh, K. Hirdesh, and K. Sanjay. *Dataset of potential targets for Mycobacterium tuberculosis H37Rv through comparative genome analysis.* Bioinformation, 4(6):245-248, 2009.

[7] World Health Organization (WHO) Report. *Global tuberculosis control: surveillance, planning, financing.* Geneva, 2008.

[8] World Health Organization (WHO) Report. *Global tuberculosis control*, 2009.

[9] S. T. Cole, R. Brosch, J. Parkhill, T. Garnier, C. M. Churcher, D. E. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. III Barry, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. M. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. D. Murphy, S. Oliver, J. Osborne, M. A. Quail, M. A. Rajandream, J. Rogers, S. Rutter, K. Seeger, S. Skelton, S. Squares, R. Squares, J. E. Sulston, K. Taylor, S. Whitehead, and B. G. Barrell. *Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence.* Nature, 393:537-544, 1998.

[10] R. D. Fleischmann, D. Alland, J. A. Eisen, L. Carpenter, O. White, J. D. Peterson, R. T. DeBoy, R. J. Dodson, M. L. Gwinn, D. H. Haft, E. K. Hickey, J. F. Kolonay, W. C. Nelson, L. A. Umayam, M. D. Ermolaeva, S. L. Salzberg, A. Delcher, T. R. Utterback, J. F. Weidman, H. M. Khouri, J. Gill, A. Mikula, W. Bishai, W. R. Jr. Jacobs, J. C. Venter, and C. M. Fraser. *Whole-genome comparison of Mycobacterium tuberculosis clinical and laboratory strains.* J. Bacteriol., 184:5479-5490, 2002.

[11] C. Dye, A. D. Harries, D. Maher, S. M. Hosseini, W. Nkhoma, and F. M. Salaniponi. *Tuberculosis-Disease and Mortality in Sub-Saharan Africa.* The World Bank at `http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=dmssa&part=A1076`, 2006.

[12] I. Vlisidou, A. J. Dowling, I. R. Evans, N. Waterfield, R. H. ffrench Constant, and W. Wood. *Drosophila Embryos as Model Systems for Monitoring Bacterial Infection in Real Time.* PLoS Pathog, 5(7): e1000518, 2009.

[13] Z. Toossi. *The inflammatory response in Mycobacterium tuberculosis infection.* Arch Immunol Ther Exp (Warsz), 48(6):513-519, 2000.

[14] J. L. Flynn and J. Chan. *Immune evasion by Mycobacterium tuberculosis: living with the enemy.* Curr Opin Immunol., 15:450-455, 2003.

[15] J. Keane, H. G. Remold, and H. Kornfeld. *Virulent Mycobacterium strains evade apoptosis of infected alveolar macrophages.* J. Immunol., 164:2016-2020, 2000.

[16] T. M. Quast and R. F. Browning. *Pathogenesis and clinical manifestations of pulmonary tuberculosis.* Dis. Mon., 52:413-419, 2006.

[17] S. H. Kaufman. *How can immunology contribute to the control of tuberculosis?* Nat. Rev. Immunol., 1:20-30, 2001.

[18] Centers for Disease Control and Prevention (CDC). *Tuberculosis (TB).* `http://www.cdc.gov/tb/topic/basics/default.htm`.

[19] K. G. Mawuenyenya, C. V. Forst, K. M. Dobos, J. T. Belisle, J. Chen, E. M. Bradbury, A. R. M. Bradbury, and X. chen. *Mycobacterium tuberculosis Functional Network Analysis by Global Subcellular Protein Profiling.* Molecular Biology of the Cell, 6:396-404, 2005.

[20] E. H. Noss, C. V. Harding, and W. H. Boom. *Mycobacterium tuberculosis inihibits MHC class II antigen procesing in mirune bone marrow macrophages.* Cell Immunol., 201:63-74, 2000.

[21] L. M. Ting, A. C. Kim, A. Cattamanchi, and J. D. Ernst. *Mycobacterium tuberculosis inihibits IFN-gamma transcriptional responses without inihibiting activation of STAT1.* J. Immunol., 163:3898-3906, 1999.

[22] K. Velmurugan, B. Chen, J. L. Miller, S. Azogue, S. Gurses, T. Hsu, M. Glickman, W. R. Jacobs, S. A. Porcelli, and V. Briken. *Mycobacterium tuberculosis nuoG Is a Virulence Gene That Inihibits Apoptosis of Infected Host Cells.* PLoS Pathog., 3:e110, 2007.

[23] E. Dubnau, P. Fontan, R. Manganelli, S. Soares-Appel, and I. Smith. *Mycobacterium tuberculosis genes induced during infection of human macrophages.* Infect Immun., 70:2787-2795, 2002.

[24] J. A. Armstrong and P. D. Hart. *Response of cultured macrophages to Mycobacterium tuberculosis, with observations on fusion of lysosomes with phagosomes.* J. Exp. Med., 134:713-740, 1971.

[25] R. A. Fratti, J. M. Backer, J. Gruenberg, S. Corvera, and V. Deretic. *Role of phosphatidylinositol 3-kinase and Rab5 effectors in phagosomal biogenesis and mycobacterial phagosome maturation arrest.* J. Cell Biol., 154:631-644, 2001.

[26] A. W. Rooyakkers and R. W. Stokes. *Absence of complement receptor 3 results in reduced binding and ingestion of Mycobacterium tuberculosis but has no significant effect on the induction of reactive oxygen and nitrogen intermediates or on the survival of the bacteria in resident and interferon-gamma activated macrophages.* Microb. Pathog., 39:57-67, 2005.

[27] P. S. Jackett, V. R. Aber, and D. B. Lowrie. *Virulence and resistance to superoxide, low pH and hydrogen peroxide among strains of Mycobacterium tuberculosis.* J. Gen. Microbiol., 104:37-45, 1978.

[28] G. Middlebrook. *Isoniazid-resistance and catalase activity of tubercle bacilli: a preliminary report.* Am. Rev. Tuberc., 69:471-472, 1954.

[29] P. G. Marjorie and R. V. Holenarasipur. *Extrapulmonary Tuberculosis: An Overview.* American Family Physician, 72(9):1761-1768, 2005.

[30] V. R. Razanamparany, D. Ménard, G. Aurégan, B. Gicquel, and S. Chanteau. *Extrapulmonary and Pulmonary Tuberculosis in Antananarivo (Madagascar).* JCM, 40(11):3964-3969, 2002.

[31] Robert Koch and Tuberculosis. *Koch's famous lecture.* `http://nobelprize.org/educational-games/medecine/tuberculosis/readmore.html`.

[32] Kofi A. Annan. *Message on World TB day.* 24 March 2005.

[33] Global Tuberculosis Institute. *A History of Tuberculosis Treatment.* `http://www.umdnj.edu/globaltb/tbhistory.htm`.

[34] K. Raman, K. Yeturu, and N. Chandra. *targetTB: A target identification pipeline for Mycobacterium tuberculosis through an interactome, reactome and genome-scale structure analysis.* BMC Systems Biology, 2:109, 2008.

[35] G. Balázsi, A. P. Heath, L. Shi, and M. L. Gennaro. *The temporal response of the Mycobacterium tuberculosis gene regulatory network during growth arrest.* Mol. Syst. Biol., 4:225, 2008.

[36] P. Chen, J. Gearhart, M. Protopopova, L. Einck, and C. A. Nacy. *Synergistic interactions of SQ109, a new ethylene diamine, with front-line antitubercular drugs in vitro.* J. of Antimicrobial Chemotherapy, 58(2):332-337, 2006.

[37] F. Brossier, N. Veziris, A. Aubry, V. Jarlier, and W. Sougakoff. *Detection by GenoType MTBDRsl Test of Complex Mechanisms of Resistance to Second-Line Drugs and Ethambutol in Multidrug-Resistant Mycobacterium tuberculosis Complex Isolates.* J. Clin. Microbiol., 48:1683-1689, 2010.

[38] P. Jureen, K. Angeby, E. Sturegard, E. Chryssanthou, C. G. Giske, J. Werngren, M. Nordvall, A. Johansson, G. Kahlmeter, S. Hoffner, and T. Schon. *Detection by Genotype MTBDRsl test of complex resistance mechanisms to second-line drugs and ethambutol in multidrug-resistant Mycobacterium tuberculosis complex isolates.* J. Clin. Microbiol., 48:1853-1858, 2010.

[39] Global Alliance for TB Drug Development: Drug-Resistant TB. *http://www.tballiance.org/why/mdr-tb.php*.

[40] C. D. Wells, J. P. Cegielski, L. J. Nelson, K. F. Laserson, T. H. Holtz, A. Finlay, K. G. Castro, and K. Weyer. *HIV infection and multidrug-resistant tuberculosis: the perfect storm.* J. Infect. Dis., 196(1):S86-107, 2007.

[41] K. J. Seung, D. B. Omatayo, S. Keshavjee, J. J. Furin, P. E. Farmer, and H. Satti. *Early Outcomes of MDR-TB Treatment in a High HIV-Prevalence Setting in Southern Africa.* PLoS ONE, 4(9): e7186, 2009.

[42] N. Mandela. *XV International AIDS Conference* *http://www.healthinitiative.org/html/conf/thai/index.htm, July2004*.

[43] World Health Organization (WHO) Report. *TB/HIV Facts: The Challenge.* Geneva, 2009.

[44] M. B. Prentice. *Bacterial comparative genomics.* Genome Biology, 5:338, 2004.

[45] J. J. Ferretti, D. Ajdic, and W. M. McShan. *Comparative genomics of streptococcal species.* Indian J. Med Res., 119(Suppl):1-6, 2004.

[46] J.-C. Camus, M. J. Pryor, C. Medigue, and S. T. Cole. *Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv.* Microbiology, 148:2967-2973, 2002.

[47] M. Daffe and P. Drapper. *The envelope layers of mycobacteria with reference to their pathogenicity.* Adv. Microb. Physiol., 39:131-203, 1998.

[48] P. Constant, E. Perez, W. Malaga, M. A. Laneelle, O. Saurel, M. Daffe, and C. Guilhot. *Role of the pks15/1 gene in biosynthesis of the phenolglycolipids in the Mycobacterium tuberculosis complex. Evidence that all strains synthesize glycosylated p-hydroxybenzoic methyl esters and that strains devoid of phenolglycolipids harbor a frameshift mutation in the pks15/1 gene.* J. Biol. Chem., 277:38148-38158, 2002.

[49] M. B. Reed, P. Domenech, C. Manca, H. Su, A. K. Barezak, B. N. Kreiswirth, G. Kaplan, and C. E. Barry 3rd. *A glycolipid of hypervirulent tuberculosis strains that inihibits the innate immune response.* Nature, 431:84-87, 2004.

[50] S. T. Cole. *Comparative and functional genomics of the Mycobacterium tuberculosis complex.* Microbiology, 148(Pt 10):2919-2928, 2002.

[51] A. M. Abdallah, T. Verboom, E. M. Weerdenburg, N. C. Gey van Pittius, P. W. Mahasha, C. Jiménez, M. Parra, N. Cadieux, M. J. Brennan, B. J. Appelmelk, and W. Bitter. *PPE and PE_PGRS proteins of Mycobacterium marinum are transported via the type VII secretion system ESX-5.* Molecular Microbiology, 73(3):329-340, 2009.

[52] G. Delogu and M. J. Brennan. *Comparative Immune Response to PE and PE_PGRS Antigens of Mycobacterium tuberculosis.* Infect. Immun., 69(9):5606-5611, 2001.

[53] M. J. Brennan, G. Delogu, Y. Chen, S. Bardarov, J. Kriakov, M. Alavi, and W. R. Jacobs Jr. *Evidence that Mycobacterial PE_PGRS Proteins Are Cell Surface Constituents That Influence Interactions with Other Cells.* Infect. Immun., 69(12):7326-7333, 2001.

[54] S. Banu, N. Honore, B. Saint-Joanis, D. Philpott, M. C. Prevost, and S. T. Cole. *Are the PE-PGRS proteins of Mycobacterium tuberculosis variable surface antigens?* Mol. Microbiol., 44(1):9-19, 2002.

[55] Y. Huang, Y. Wang, Y. Bai, Z. G. Wang, L. Yang, and D. Zhao. *Expression of PE_PGRS 62 protein in Mycobacterium smegmatis decrease mRNA expression of proinflammatory cytokines IL-1β, IL-6 in macrophages.* Mol. Cell. Biochem., 2010.

[56] K. K. Singh, X. Zhang, A. S. Patibandla, P. Chien Jr., and S. Laal. *Antigens of Mycobacterium tuberculosis Expressed during Preclinical Tuberculosis: Serological Immunodominance of Proteins with Repetitive Amino Acid Sequences.* Infect. Immun., 69(6):4185-4191, 2001.

[57] M. J. Brennan and G. Delogu. *The PE multigene family: a 'molecular mantra' for mycobacteria.* Trends in Microbiol., 10(5):246-249, 2002.

[58] A. Karboul, A. Mazza, N. C. Gey van Pittius, J. L. Ho, R. Brousseau, and H. Mardassi. *Frequent Homologous Recombination Events in Mycobacterium tuberculosis PE/PPE Multigene Families: Potential Role in Antigenic Variability.* Journal of Bacteriology, 190(23):78387846, 2008.

[59] M. I. Voskuil, D. Schnappinger, R. Rutherford, Y. Liu, and G. K. Schoolnik. *Regulation of the Mycobacterium tuberculosis PE/PPE genes.* Tuberculosis (Edinb), 84(3-4):256-62, 2004.

[60] S. E. Vasconcellos, R. C. Huard, S. Niemann, K. Kremer, A. R. Santos, P. N. Suffys, and J. L. Ho. *Distinct genotypic profiles of the two major clades of Mycobacterium africanum.* BMC Infect Dis, 10:80, 2010.

[61] P. Brodin, K. Eiglmeier, M. Marmiesse, A. Billault, T. Garnier, S. Niemann, S. T. Cole, and R. Brosch. *Bacterial artificial chromosome-based comparative genomic analysis identifies Mycobacterium microti as a natural ESAT-6 deletion mutant.* Infect Immun., 70(10):5568-5578, 2002.

[62] T. Hsu, S. M. Hingley-Wilson, B. Chen, M. Chen, A. Z. Dai, P. M. Morin, C. B. Marks, J. Padiyar, C. Goulding, M. Gingery, D. Eisenberg, R. G. Russell, S. C. Derrick, F. M. Collins, S. L. Morris, C. H. King, and Jr. W. R. Jacobs. *The primary mechanism of attenuation of bacillus Calmette-Guérin is a loss of secreted lytic function required for invasion of lung interstitial tissue.* PNAS, 100(21):12420-12425, 2003.

[63] C. S. Aagaard, T. T. Hoang, C. Vingsbo-Lundberg, J. Dietrich, and P. Andersen. *Quality and vaccine efficacy of CD4+ T cell responses directed to dominant and subdominant epitopes in ESAT-6 from Mycobacterium tuberculosis.* J. Immunol., 183(4):2659-68, 2009.

[64] A. S. Pym, P. Brodin, R. Brosch, M. Huerre, and S. T. Cole. *Loss of RD1 contributed to the attenuation of the live tuberculosis vaccines Mycobacterium bovis BCG and Mycobacterium microti.* Mol. Microbiol., 46(3):709-717, 2002.

[65] I. M. Orme. *The search for new vaccines against tuberculosis.* Journal of Leukocyte Biology, 70:1-10, 2001.

[66] R. Brosch, W. J. Philipp, E. Stavropoulos, M. J. Colston, S. T. Cole, and S. V. Gordon. *Genomic analysis reveals variation between Mycobacterium tuberculosis H37Rv and the attenuated M. tuberculosis H37Ra strain.* Infect. Immun., 67(11):5768-5774, 1999.

[67] A. S. Mustafa. *Vaccine potential of Mycobacterium tuberculosis-specific genomic regions: in vitro studies in humans.* Expert Rev. Vaccines, 8(10):1309-1312, 2009.

[68] R. Brosch, S. V. Gordon, M. Marmiesse, P. Brodin, C. Buchrieser, K. Eiglmeier, T. Garnier, C. Gutierrez, G. Hewinson, K. Kremer, L. M. Parsons, A. S. Pym, S. Samper, D. van Soolingen, and S. T. Cole. *A new evolutionary scenario for the Mycobacterium tuberculosis complex.* PNAS, 99(6):3684-3689, 2002.

[69] Z. Fang, C. Doig, D. T. Kenna, N. Smittipat, P. Palittapongarnpim, B. Watt, and K. J. Forbes. *IS6110-mediated deletions of wild-type chromosomes of Mycobacterium tuberculosis.* J Bacteriol., 181(3):1014-1020, 1999.

[70] A. G. Tsolaki, A. E. Hirsh, K. DeRiemer, J. A. Enciso, M. Z. Wong, M. Hannan, Y.-O. L. Goguet de la Salmoniere, K. Aman, M. Kato-Maeda, and P. M. Small. *Functional and evolutionary genomics of Mycobacterium tuberculosis: Insights from genomic deletions in 100 strains.* PNAS, 101(14):4865-4870, 2004.

[71] S. Mostowy, C. Cletoa, D. R. Shermanb, and M. A. Behr. *The Mycobacterium tuberculosis complex transcriptome of attenuation.* Tuberculosis, 84(3):131-137, 2004.

[72] Oxford Immunotec. *Harnessing the power of T-cell measurement.* `http://www.oxfordimmunotec.com`.

[73] Healthier You. *Tuberculosis,* ***http://www.healthieryou.com/tb.html***.

[74] Tuberculosis: Prevention.
http://www.answers.com/topic/tuberculosis-prevention.

[75] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. *A basic local alignment search tool*. Journal of Molecular Biolology, 215(3):403-410, 1990.

[76] B. Potter, K. Rindfleisch, and C. K. Kraus. *Management of Active Tuberculosis*. American Family Physician, 72:2225-2232, 2005.

[77] E. Amukoye. *Multi drug resistant tuberculosis: a challenge in the management of tuberculosis*. African Journal of Health Sciences, 15(1):6-13, 2008.

[78] F. Browne, H. Zheng, H. Wang, and F. Azuaje. *An Integrative Bayesian Approach to Supporting the Prediction of Protein-Protein Interactions: A Case Study in Human Heart Failure*. World Academy of Science, Engineering and Technology, 53:457-463, 2009.

[79] K. Raman and N. Chandra. *Mycobacterium tuberculosis interactome analysis unravels potential pathways to drug resistance*. BMC Microbiology, 8:234, 2008.

[80] M. Chagoyen and F. Pazos. *Quantifying the biological significance of gene ontology biological processes-implications for the analysis of systems-wide data*. Bioinformatics, 26(3):378-384, 2010.

[81] C. von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen, and P. Bork. *STRING: known and predicted protein-protein associations, integrated and transferred across organisms*. Nucleic Acids Research, 33:D433-D437, 2005.

[82] L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering. *STRING 8-a global view on proteins and their functional interactions in 630 organisms*. Nucleic Acids Res., 37:D412-D416, 2008.

[83] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. *KEGG: Kyoto Encyclopedia of Genes and Genomes*. ucleic Acids Res., 27(1):29-34, 1999.

[84] L. J. Jensen, J. Lagarde, C. von Mering, and P. Bork. *ArrayProspector: a web resource of functional associations inferred from microarray expression data.* Nucleic Acids Research, 32:W445-W448, March 2004.

[85] N. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, V. Buillard, L. Cerutti, R. Copley, E. Courcelle, U. Das, L. Daugherty, M. Dibley, R. Finn, W. Fleischmann, J. Gough, D. Haft, N. Hulo, S. Hunter, D. Kahn, A. Kanapin, A. Kejariwal, A. Labarga, P. S. Langendijk-Genevaux, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, A. N. Nikolskaya, S. Orchard, C. Orengo, R. Petryszak, J. D. Selengutand C. J. Sigrist, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, and C. Yeats. *New developments in the InterPro database.* Nucleic Acids Res., 35:D224-D228, 2007.

[86] M. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bradley, P. Bork, P. Bucher, L. Cerutti, R. Copley, E. Courcelle, U. Das, R. Durbin, W. Fleischmann, J. Gough, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, J. McDowall, A. Mitchell, A. N. Nikolskaya, S. Orchard, M. Pagni, C. P. Ponting, E. Quevillon, J. Selengut, C. J. Sigrist, V. Silventoinen, D. J. Studholme, R. Vaughan, and C. H. Wu. *InterPro, progress and status in 2005.* Nucleic Acids Res., 33:D201-D205, 2005.

[87] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. *Gene Ontology: tool for the unification of biology.* Nature Genetics, 25(1):25-29, 2000.

[88] C. L. Myers and O. G. Troyanskaya. *Context data integration and prediction of biological networks.* Bioinformatics, 23(17):2322-2330, 2007.

[89] H. N. Chua, W. K. Sung, and L. Wong. *An efficient strategy for extensive integration of diverse biological data for protein function prediction.* Bioinformatics, 23(24):3364-3373, 2007.

[90] M. A. Mahdavi and Y.-H. Lin. *Prediction of Protein-Protein Interactions Using Protein Signature Profiling.* Genomics, Proteomics & Bioinformatics, 5(3-4):177-186, 2007.

[91] X. Mao, T. Cai, J. G. Olyarchuk, and L. Wei. *Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary.* Bioinformatics, 21(19):3787-3793, 2005.

[92] S. Yellaboina, K. Goyal, and S. C. Mande. *Inferring genome-wide functional linkages in E. coli by combining improved genome context methods: Comparison with high-throughput experimental data.* Genome Research, 17:527-535, 2007.

[93] J. Krawczyk, T. A. Kohl, A. Goesmann, J. Kalinowski, and J. Baumbach. *From Corynebacterium glutamicum to Mycobacterium tuberculosis-towards transfers of gene regulatory network and integrated data analyses with MycoRegNet.* Nucleic Acids Res., 37(14):e97, 2009.

[94] O. Bastian, P. Ortet, S. Roy, and E. Maréchal. *A configuration space of homologous proteins conserving mutual information and allowing a phylogeny inference based on pair-wise Z-score probabilities.* BMC Bioinformatics, 6:49, 2005.

[95] O. Bastian and E. Maréchal. *Evolution of Biological sequences implies an extrema value distribution of type I for both global and local pairwise alignments scores.* BMC Bioinformatics, 9:332, August 2008.

[96] R. V. L. Hartley. *Transmission of Information.* The Bell System Technical Journal, 3:535-564, 1928.

[97] C. E. Shannon. *A Mathematical Theory of Communication.* The Bell System Technical Journal, 27:379-423, 1948.

[98] NCBI. *The Statistics of Sequence Similarity Scores.* `http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html`.

[99] W. R. Pearson. *Protein sequence comparison and Protein evolution.* Tutorial-ISBM2000, October 2001.

[100] David J. C. Mackay. *Information Theory, Inference, and Learning algorithms.* Cambridge University Press, August 2004.

[101] S. F. Altschul. *Amino acid substitution matrices from an information theoretic perspective.* J. Mol. Biol., 219:555-565, 1991.

[102] G. Subramanian, K. V. Koonin, and L. Aravind. *Comparative Genome Analysis of the Pathogenic Spirochetes Borrelia burgdorferi and Treponema pallidum.* Infection and Immunity, 68(3):1633-1648, 2000.

[103] P. G. Aaron, M. L. Sonia, A. B. William, Lawrence E. H., and S. G Debra. *Improving protein function prediction methods with integrated literature data.* BMC Bioinformatics, 9:198, 2008.

[104] C. von Mering, L. J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Krüger, B. Snel, and P. Bork. *STRING 7-recent developments in the integration and prediction of protein interactions.* Nucleic Acids Research, 35:D358-D362, 2007.

[105] F. Crick. *On Protein Synthesis.* Symp. Soc. Exp. Biol. XII, 139-163, 1958.

[106] F. Crick. *Central Dogma of Molecular Biology.* Nature, 227:561-563, 1970.

[107] S. DraGhici. *Data Analysis tools for DNA Microarrays.* Chapman & Hall / CRC Mathematical Biology and Medecine Series, ISBN: 1-58488-315-4, 2003.

[108] X. Gao, D. Q. Pu, and P. X.-K. Song. *Transition Dependency: A Gene-Gene Interaction Measure for Times Series Microarray Data.* EURASIP Journal on Bioinformatics and Systems Biology, Vol 2009, Nov. 2008.

[109] S. Datta. *Exploring relationships in gene expressions: A partial least square approach.* Gene Expression, 9: 257-264, 2001.

[110] V. Pihur, Somnath Datta, and Susmita Datta. *Reconstruction of genetic association networks from microarray data: A partial least square approach.* Systems biology, 24(4):561-568, Jan. 2008.

[111] P. Langfelder and S. Horvath. *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 9:559, 2008.

[112] A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Della Favera, and A. Califano. *Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.* BMC Bioinformatics, 7(1):S7, 2006.

[113] F. Markowetz and R. Spang. *Inferring Cellular networks-a review.* BMC Bioinformatics, 8(6):S7, 2007.

[114] J. Schäfer and K. Strimmer. *An Empirical Bayes Approach to Inferring Large-Scale Gene Association Networks.* Bioinformatics, 21:754-764, 2005.

[115] R. Opgen-Rhein and K. Strimmer. *From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data.* BMC Systems Biology, 1:37, 2007.

[116] K. Basso, A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano. *Reverse engineering of regulatory networks in human B cells.* Nat. Genet., 37:382-390, 2005.

[117] J. Yu, V. A. Smith, P. P. Wang, A. J. Hartemink, and E. D. Jarvis. *Advances to bayesian network inference for generating causal networks from observational biological data.* Bioinformatics, 20:3594-3603, 2004.

[118] W. Wu and R. Manne. *Fast regression methods in a Lanczos (or PLS-1) basis: Theory and applications.* Chemometrics and Intelligent Laboratory Systems, 51:145-161, 2000.

[119] B. S. Dayal and J. F. MacGregor. *Improved PLS algorithms.* Journal of Chemometrics, 11:73-85, 1997.

[120] S. Wold, M. Sjöström, and L. Ericksson. *PLS-regression: A basic tool of chemometrics.* Chemometrics and Intelligent Laboratory Systems, 58:109-130, 2001.

[121] Z. Liu and D. Chen. *Gene Expression Data Classification with Revised Kernel Partial Least Squares algorithm.* American Association for Artificial Intelligence, 2004.

[122] R. Li, G. Meng, N. Gao, and H. Xie. *Combined use of partial least-squares regression and neural network for residual life estimation of large generator stator insulation.* Measurement science and Technology, 18:2074-2082, 2007.

[123] H. Abdi. *Partial Least Squares Regression and Projection on Latent Structure Regression (PLS-Regression).* Wiley Interdisciplinary Reviews: Computational Statistics, Jan. 2009.

[124] S. Maitra and J. Yan. *Principle Component Analysis and Partial Least Squares: Two Dimension Reduction Techniques for Regression.* Causality Actuarial Society, 2008.

[125] T. Barrett, T. O. Suzek, D. B. Troup, S. E. Wilhite, W.-C. Ngau, P. Ledoux, D. Rudnev, A. E. Lash Wataru Fujibuchi, and R. Edgar. *NCBI GEO: mining millions of expression profiles-database and tools.* Nucleic Acids Research, 33:D562-D566, 2005.

[126] G. Sherlock, T. Hernandez-Boussard, A. Kasarskis, G. Binkley, J. C. Matese, S. S. Dwight, M. Kaloper, S. Weng, H. Jin, C. A. Ball, M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, and J. M. Cherry. *The Stanford Microarray Database.* Nucleic Acids Research, 29(1):152-155, 2001.

[127] A. Höskuldsson. *PLS regression methods.* Journal of Chemometrics, 2:211-228, 1988.

[128] R. Rosipal and L. J. Trejo. *Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space.* Journal of Machine Learning Research, 2:97-123, 2001.

[129] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis.* Cambridge Univesity Press, ISBN: 0-52146-713-6, June 1994.

[130] D. V. Nguyen and D. M. Rocke. *Tumor classification by partial least squares using microarray gene expression data.* Bioinformatics, 18(1):39-50, 2002.

[131] M. I. Griep, I. N. Wakeling, P. Vankeerberghen, and D. L. Massart. *Comparison of semirobust and robust partial least squares procedures.* Chemometrics and Intelligent Laboratory System, 29:37-50, 1995.

[132] A.J. Burnham, R. Viveros, and J.F. MacGregor. *Frameworks for latent variable multivariate regression.* Journal of Chemometrics, 10:31-45, 1996.

[133] I. N. Wakeling and J. J. Morris. *A test of significance for partial least squares regression.* Journal Chemometrics, 7:291-304, 1993.

[134] M. C. Denham. *Prediction Intervals in Partial Least Squares.* Journal of Chemometrics, 11:39-52, 1997.

[135] X. Huang, W. Pan, S. Park, X. Han, Y. Chen, L. W. Miller, and J. L. Hall. *Modeling the relationship between LVAD support time and gene expression changes in the human heart by penalized partial least squares.* Bioinformatics, 20(6):888-894, 2004.

[136] X. Huang and W. Pan. *Linear regression and two-class classification with gene expression data.* Bioinformatics, 19:2072-2078, 2003.

[137] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap.* Chapman & Hall, ISBN: 0-412-04231-0, 1993.

[138] S. Sahinler and D. Topuz. *Bootstrap and Jackknife Resampling Algorithms for Estimation of Regression Parameters.* Journal of Applied Quantitative Methods, 2:188-199, 2007.

[139] K. E. Kramer, R. E. Morris, S. L. Rose-Pehrsson, J. Cramer, and K. J. Johnson. *Statistical Significance Testing as a Guide to Partial Least-Squares (PLS) Modeling of Nonideal Data Sets for Fuel Property Predictions.* Energy Fuels, 22(1):523-534, 2008.

[140] G. Blanchard and N. Krämer. *Kernel Partial Least Squares is Universally Consistent.* Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS), 2010.

[141] N. Krämer and M. L. Braun. *plsdof - Degrees of Freedom for Partial Least Squares Regression.* R package version 0.2-0, `http://cran.at.r-project.org/web/packages/plsdof/index.html`, 2010.

[142] A. M. Lesk. *Introduction to Bioinformatics.* Oxford University Press, Second Edition ISBN 0-19-927787-7, 2005.

[143] J. Gagneur, R. Krause, T. Bouwmeester, and G. Casari. *Modular decomposition of protein-protein interaction networks.* Genome Biology, 5:R57, 2004.

[144] C. Lee and M.-H. Yu. *Protein Folding and Diseases.* Journal of Biochemistry and Molecular Biology, 38(3):275-280, 2005.

[145] M. Kumar and G. P. S. Raghava. *Prediction of nuclear proteins using SVM and HMM models.* BMC Bioinformatics, 10:22, 2009.

[146] A. Szilágyi, V. Grimm, A. K. Arakaki, and J. Skolnick. *Prediction of physical protein-protein interactions.* Physical Biology, 2:S1-S16, 2005.

[147] E. M. Marcotte, M. Pellegrini, H.-L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. *Detecting protein function and protein-protein interactions from genome sequences.* Science, 285(5428):751-753, 1999.

[148] D. Frishman and A. Valencia. *Modern Genome Annotation.* Spring-Verlag, ISBN: 978-3-211-75122-0, 2009.

[149] R. Caspi, T. Altman, J. M. Dale, K. Dreher, C. A. Fulcher, F. Gilham, P. Kaipa, A. S. Karthikeyan, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, S. Paley, L. Popescu, A. Pujar, A. G. Shearer, P. Zhang, and P. D. Karp. *The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases.* Nucleic Acids Res., D473-D479, 2010.

[150] M . Pruess, P. Kersey, and R. Apweiler. *The Integr8 project–a resource for genomic and proteomic data.* In Silico Biol., 5(2):179-185, 2004.

[151] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh. *UniProt: the Universal Protein knowledgebase.* Nucleic Acids Res., 32:D115-D119, 2004.

[152] T. Dandekar, B. Snel, M. Huynen, and P. Bork. *Conservation of gene order: a fingerprint of proteins that physically interact.* Trends Biochem Sci., 23(9):324-328, 1998.

[153] H. Salgado, G. Moreno-Hagelsieb, T. F. Smith, and J. Collado-Vides. *Operons in Escherichia coli: genomic analyses and predictions.* PNAS, 97(12):6652-6657, 2000.

[154] M. N. Price, A. P. Arkin, and E. J. Alm. *The Life-Cycle of Operons.* PLoS Genet., 2(6):e96, 2006.

[155] T. Gaasterland and M. A. Ragan. *Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes.* Microb. Comp. Genomics,3(4):199-217, 1998.

[156] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. *Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles.* PNAS, 96:4285-4288, 1999.

[157] S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau. *Text Mining: Predictive Methods for Analyzing Unstructured Information.* Springer, 2005.

[158] N. Uramoto, H. Matsuzawa, T. Nagano, A. Murakami, H. Takeuchi, and K. Takeda. *A text-mining system for knowledge discovery from biomedical documents.* IBM Systems Journal, 43(3):516-533, 2003.

[159] I. Xenarios, L. Salwínski, X. J. Duan, P. Higney, S.-M. Kim, and D. Eisenberg. *DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions.* Nucleic Acids Research, 30(1): 303-305, 2002.

[160] O. Mason and M. Verwoerd. *Graph theory and networks in Biology.* IET Syst Biol., 1(2):89-119, 2007.

[161] A. Gursoy, O. Keskin, and R. Nussinov. *Topological properties of protein interaction networks from structural perspective.* Biochem. Soc. Trans., 36:1398-1403, 2008.

[162] J. Persener. *Bioinformatics and Functional Genomics.* John Wiley & Sons, ISBN: 0-471-21004-8, 2003.

[163] F. Enault, K. Suhre, and J.-M. Claverie. *Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis.* BMC Bioinformatics, 6:24, 2005.

[164] P. Baldi and S. Brunak. *BIOINFORMATICS: The Machine Learning Approach.* Massachusetts Institute of Technology, ISBN:0-262-02506-X, 2^{nd}*Edition*, 2001.

[165] R. D. King, A. Karwath, A. Clare, and L. Dehaspe. *Accurate prediction of protein functional class from sequence in the Mycobacterium tuberculosis and Escherichia coli genomes using data mining.* Yeast, 17:283-293, 2000.

[166] J. A. Eisen. *Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis.* Genome Research, 8:163-167, 1998.

[167] P. Bork and E. V. Koonin. *Predicting functions from protein sequences-where are the bottlenecks?* Nature Genetics, 18:313-318, 1998.

[168] M. Y. Galperin and E. V. Koonin. *Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption.* In Silico Biology, 1(1):55-67, 1998.

[169] D. Devos and A. Valencia. *Practical limits of function prediction.* PROTEINS: Structure, Function, and Genetics, 41(1):98-107, 2000.

[170] M. Y. Galperin and E. V. Koonin. *Who's your neighbor? New computational approaches for functional genomics.* Nature Biotechnology, 18:609-613, 2000.

[171] P. Bork, T. Dandekar, Y. Diaz-Lazcoz, F. Eisenhaber, M. Huynen, and Y. Yuan. *Predicting Function: From Genes to Genomes and Back.* J. Mol. Biol., 283(4):707-725, 1998.

[172] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. *Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation.* Bioinformatics, 19(10):1275-1283, 2003.

[173] T. R. Gruber. *Toward Principles for the Design of Ontologies Used for Knowledge Sharing.* International Journal Human-Computer Studies, 43(4-5):907-928, 1995.

[174] T R Gruber. *A Treanslation Approach to Portable Ontology Specifications.* Knowledge Acquisition, 5(2):199-220, 1993.

[175] R. Stevens, C. A. Goble, and S. Bechhofer. *Ontology-based knowledge representation for bioinformatics.* Briefings in Bioinformatics, 1(4):398-414, 2000.

[176] M. Ciocoiu, M. Gruninger, and D. Nau. *Ontologies for integrating engineering applications.* Journal of Computing and Information Science in Engineering, 1:45-60, 2001.

[177] M. Uschold and M. Gruninger. *Ontologies and Semantics for Seamless Connectivity.* SIGMOD Record, 33:58-64, 2004.

[178] Enzyme Nomenclature. *Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes by the Reactions they Catalyse.* `http://www.chem.qmul.ac.uk/iubmb/enzyme/`.

[179] S. C. G. Rison, T. C. Hodgman, and J. M. Thornton. *Comparison of functional annotation schemes for genomes.* Funct Integr Genomics, 1(1):56-69, 2000.

[180] C. A. Ouzounis, R. M. Coulson, A. J. Enright, V. Kunin, and J. B. Pereira-Leal. *Classification schemes for protein structure and function.* Nature Reviews Genetics, 4(7):508-519, 2003.

[181] P. D. Karp, M. Riley, M. Saier, I. T. Paulsen, J. Collado-Vides, S. M. Paley, A. Pellegrini-Toole, C. Bonavides, and S. Gama-Castro. *The EcoCyc Database.* Nucleic Acids Research, 30(1):56-58, 2002.

[182] I. M. Keseler, J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I. T. Paulsen, M. Peralta-Gil, and P. D. Karp. *EcoCyc: a comprehensive database resource for Escherichia coli.* Nucleic Acids Research, 33:D334-D337, 2005.

[183] I. M. Keseler, C. Bonavides-Martínez, J. Collado-Vides, S. Gama-Castro, R. P. Gunsalus, D. A. Johnson, M. Krummenacker, L. M. Nolan, S. Paley, I. T. Paulsen, M. Peralta-Gil, A. Santos-Zavaleta, A. G. Shearer, and P. D. Karp. *EcoCyc: A comprehensive view of Escherichia coli biology.* Nucleic Acids Research, 37:D464-D470, 2009.

[184] Q. Zheng and X.-J. Wang. *GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis.* Nucleic Acids Research, 36(2):W358-W363, 2008.

[185] The Gene Ontology. *Ontology Structure.* `http://www.geneontology.org/GO.ontology.structure.shtml`.

[186] The Gene Ontology. *Ontology relations.* `http://www.geneontology.org/GO.ontology.relations.shtml`.

[187] The Gene Ontology. *The GO Flat File Format.* `http://www.geneontology.org/GO.format.go.shtml`.

[188] GO Consortium. *The Gene Ontology in 2010: extensions and refinements.* 2009.

[189] GO Consortium. *The Gene Ontology (GO) project in 2006.* 2006.

[190] S. Carbon, A. Ireland, C. J. Mungall, S. Shu, B. Marshall, S. Lewis, The AmiGO Hub, and the Web Presence Working Group. *AmiGO: online access to ontology and annotation data.* Bioinformatics 25(2):288-289, 2009.

[191] C. Pesquita, D. Faria, A. O. Falcão, P. Lord, and F. M. Couto. *Semantic Similarity in Biomedical Ontologies.* PLoS Comput Biol, 5(7):e1000443, 2009.

[192] E. Camon, M. Magrane, D. Barrell, D. Binns, W. Fleischmann, P. Kersey, N. Mulder, T. Oinn, J. Maslen, A. Cox, and R. Apweiler. *The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro.* Genome Research, 13(4):662-672, 2003.

[193] E. Camon, D. Barrell, V. Lee, E. Dimmer, and R. Apweiler. *The Gene Ontology Annotation (GOA) Database - An integrated resource of GO annotations to the UniProt Knowledgebase.* In Silico Biology 4, 2003.

[194] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler. *The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology.* Nucleic Acids Research, 32:D262-D266, 2004.

[195] D. Barrell, E. Dimmer, R. P. Huntley, D. Binns, C. O'Donovan, and R. Apweiler. *The GOA database in 2009-an integrated Gene Ontology Annotation resource.* Nucleic Acids Research, 37:D396-D403, 2009.

[196] GOA FAQ. *What is GOA?,* `http://wiki.geneontology.org/index.php/GOA_FAQ`.

[197] E. C. Dimmer, R. P. Huntley, D. G. Barrell, D. Binns, S. Draghici, E. B. Camon, M. Hubank, P. J. Talmud, R. Apweiler, and R. C. Lovering. *The Gene Ontology - Providing a Functional Role in Proteomic Studies.* Proteomics, 8(Suppl), 2008.

[198] The Gene Ontology. *Guide to GO Evidence Codes.* `http://www.geneontology.org/GO.evidence.shtml`.

[199] L. N. Soldatova and R. D. King. *Are the current ontologies in biology good ontologies?* Nat Biotech, 24:902-903, 2005.

[200] J. Shon, John Y. Park, and L. Wei. *Beyond similarity-based methods to associate genes for the inference of function.* Biosilico, 1(3):89-96, 2003.

[201] F. Shi, Q. Chen, and X. Niu. *Functional Similarity Analyzing of Protein Sequences with Empirical Mode Decomposition.* Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007), 2:766-770, 2007.

[202] T. Kambe, T. Suzuki, M. Nagao, and Y. Yamaguchi-Iwai. *Sequence Similarity and Functional Relationship Among Eukaryotic ZIP and CDF Transporters.* Genomics, Proteomics & Bioinformatics, 4(1):1-9, 2006.

[203] J. L. Sevilla, V. Segura, A. Podhorski, E. Guruceaga, J. M. Mato, L. A. Martinez-Cruz, and F. J. Corrales. *Correlation between Gene Expression and GO Semantic Similarity.* IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) archive, 2(4):330-338, 2005.

[204] T. J. Hestilow and Y. Huang. *Clustering of Gene Expression Data Based on Shape Similarity.* EURASIP Journal on Bioinformatics and Systems Biology, 2009:12-pages, 2009.

[205] W. Wang, M. J. Cherry, Y. Nochomovitz, E. Jolly, D. Botstein, and H. Li. *Inference of combinatorial regulation in yeast transcriptional networks: A case study of sporulation.* PNAS, 102(6):1998-2003, 2004.

[206] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wand. *GOSemSim: an R package for measuring semantic similarity among GO terms and gene products.* Bioinformatics, 26(7):976-978, 2010.

[207] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen. *A new method to measure the semantic similarity of GO terms.* Bioinformatics, 23(10):1274-1281, 2007.

[208] P. Resnik. *Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language.* Journal of Artificial Intelligence Research, 11:95-130, 1999.

[209] D. Lin. *An Information-Theoretic Definition of Similarity.* Proceedings of the Fifteenth International Conference on Machine Learning: 296-304, 1998, 1998.

[210] A. Schlicker, F. S. Domingues, J. Rahnenführer, and T. Lengauer. *A new measure for functional similarity of gene products based on Gene Ontology.* BMC Bioinformatics, 7:302, 2006.

[211] D. Martin, C. Brun, E. Remy, P. Mouren, D. Thieffry, and B. Jacq. *GOToolBox: Functional Analysis of Gene Data Sets Based on Gene Ontology.* Genome Biology, 5(12):1901-1908, 2004.

[212] I. Friedberg. *Automated protein function prediction-the genomic challenge.* Briefings in Bioinformatics, 7(3):225-242, 2006.

[213] A. Vazquez, A. Flammini, A. Maritan, and Vespignani A. *Global protein function prediction from protein-protein interaction networks.* Nature Biotechnology, 21(6):697-700, 2003.

[214] K. Tsuda, H. Shin, and B. Schölkopf. *Fast protein classification with multiple networks.* Bioinformatics, 21:ii59-ii65, 2005.

[215] E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh. *Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps.* Bioinformatics, 21(1):i302-i310, 2005.

[216] O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altman, and D. Botstein. *A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae).* PNAS, 100(14):8348-8353, 2003.

[217] M. Deng, T. Chen, and F. Sun. *An Integrated Probabilistic Model for Functional Prediction of Proteins.* Journal of Computational Biology, 11(2-3):463-475, 2004.

[218] S. Letovsky and S. Kasif. *Predicting protein function from protein/protein interaction data: a probabilistic approach.* Bioinformatics, 19(Suppl 1):i197-i204, 2003.

[219] Y.-R. Cho, L. Shi, M. Ramanathan, and Zhang A. *A probabilistic framework to predict protein function from interaction data integrated with semantic knowledge.* BMC Bioinformatics, 9:382, 2008.

[220] G. R. G. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble. *Kernel-Based Data Fusion and Its Application to Protein Function Prediction in Yeast.* Pacific Symposium on Biocomputing, 9:300-311, 2004.

[221] Y. Chen and D. Xu. *Global protein function annotation through mining genome-scale data in yeast Saccharomyces cerevisiae.* Nucleic Acids Research, 32(21):6414-6424, 2004.

[222] J. Xiong, S. Rayner, K. Luo, Y. Li, and S. Chen. *Genome wide prediction of protein function via a generic knowledge discovery approach based on evidence integration.* BMC Bioinformatics, 7:268, 2006.

[223] B. Schwikowski, P. Uetz, and S. Fields. *A network of proteinprotein interactions in yeast.* Nature Biotechnology, 18(12):1257-1261, 2000.

[224] T. M. Murali, C. J. Wu, and S. Kasif. *The art of gene function prediction.* Nature Biotechnology, 24(12):1474-1475, 2006.

[225] H.-J. Jin and H.-G. Cho. *Computational Method for Protein Function Prediction by Constructing Protein Interaction Network Dictionary.* International Journal of Pattern Recognition and Artificial Intelligence, 20(2):285-295, 2006.

[226] H. N. Chua, W. K. Sung, and L. Wong. *Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions.* Bioinformatics, 22:1623-1630, 2006.

[227] H. N. Chua, W. K. Sung, and L. Wong. *Using Indirect Protein Interactions for the Prediction of Gene Ontology Functions.* BMC Bioinformatics, 8(4):S8, 2007.

[228] H. Hishigaki, K. Nakai, T. Ono, A. Tanigami, and T. Takagi. *Assessment of prediction accuracy of protein function from protein-protein interaction data.* Yeast, 18(6):523-31, 2001.

[229] M. Deng, F. Sun, and T. Chen. *Assessment of the reliability of protein-protein interactions and protein function prediction.* Pacific Symposium on Biocomputing, 8:140-151, 2003.

[230] X. Jiang, N. Nariai, M. Steffen, S. Kasif, and E. D. Kolaczyk. *Integration of relational and hierarchical network information for protein function prediction.* BMC Bioinformatics, 9:350, 2008.

[231] J. Swets. *Measuring the accuracy of diagnostic systems.* Science, 240:1285-1293, 1988.

[232] J. Swets, R. Dawes, and J. Monahan. *Better decisions through science.* Scientific American, 283(4):82-87, 2000.

[233] M. Buckland and F. Gey. *The relationship between recall and precision.* Journal of the American Society for Information science, 45:12-19, 1994.

[234] J. Rosamond and A. Allsop. *Harnessing the Power of the Genome in the Search for New Antibiotics.* Science, 287:1973-1976, 2000.

[235] M. G. Chaitra, M. S. Shaila, and R. Nayak. *Characterization of T-cell immunogenicity of two PE/PPE proteins of Mycobacterium tuberculosis.* J. of Medical Microbiology, 57:1079-1086, 2008.

[236] N. C. Gey van Pittius, S. L. Sampson, H. Lee, Y. Kim, P. D. van Heiden, and R. M. Warren. *Evolution and expansion of the Mycobacterium tuberculosis PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions.* BMC Evolutionary Biology, 6:95, 2006.

[237] B. Ruhnau. *Eigenvector-centrality a node-centrality?* Social Networks, 22:357-365, 2000.

[238] L. C. Freeman. *A set of measures of centrality based on betweenness.* Sociometry, 40(1):35-41, 1977.

[239] A Hagberg, D. Schult, and P. Swart. *NetworkX Reference.* Release 1.1, 2010.

[240] P. Bonacich. *Some unique properties of eigenvector centrality.* Social Networks, 29:555-564, 2007.

[241] C. M. Sassetti, D. H. Boyd, and E. J. Rubin. *Genes required for mycobacterial growth defined by high density mutagenesis.* Molecular Microbiology, 48(1):77-84, 2003.

[242] C. M. Sassetti and E. J. Rubin. *Genetic requirements for mycobacterial survival during infection.* PNAS, 100(22):12989-12994, 2003.

VITA

Gaston KUZAMUNU MAZANDU was born in Kinshasa, Democratic Republic of Congo (DRC) former Zaïre on August 18, married to *Malungidi Mbambi Marie Paul*, and father of four children *Jemima Kuzamunu Mambote, Glodi Kuzamunu Mazandu, Keren Kuzamunu Kinzuemi and Emmanuel Kuzamunu Malungidi*. He graduated with BSc Honours in Mathematics with a focus on Computer Science from University of Kinshasa in 1996. He was teacher of Mathematics at secondary school (1996-1999), teaching assistant at University of Kinshasa (1999-2004) and at Catholic Faculties of Kinshasa (2002-2004). He was admitted to study at the African Institute for Mathematical Sciences (AIMS), Cape Town, South Africa on September, 2004 where he obtained a postgraduate diploma in Mathematical Sciences in June 2005. He earned his MSc degree in Computer Science with a focus on network routing algorithms from University of Stellenbosch, South Africa in December 2007. He entered University of Cape Town, South Africa in February, 2008 and joined **Prof. Nicola J. Mulder**'s Computational Biology (CBIO) research group. He concluded his research at University of Cape Town in August 2010 with a focus on Mycobacterium tuberculosis genome analysis and submitted his thesis for a PhD degree in Computational Biology.

This thesis was typeset with LaTeX by the author.

E-mail address: gmazandu@cbio.uct.ac.za